Journal of Machine Learning Research 14 (2013) xx-yy Submitted 7/13; Published ??/??

Noisy Sparse Subspace Clustering

Yu-Xiang Wang YUXIANGW@CS.CMU.EDU
Machine Learning Department
Carnegie Mellon University
Pittsburgh, PA 15213

Huan Xu MPEXUH@NUS.EDU.SG
Department of Mechanical Engineering
National University of Singapore
Singapore 117576

Editor: editors not assigned yet

Abstract

This paper considers the problem of subspace clustering under noise. Specifically, we study the behavior of Sparse Subspace Clustering (SSC) when either adversarial or random noise is added to the unlabelled input data points, which are assumed to be in a union of low-dimensional subspaces. We show that a modified version of SSC is *provably effective* in correctly identifying the underlying subspaces, even with noisy data. This extends theoretical guarantee of this algorithm to more practical settings and provides justification to the success of SSC in a class of real applications.

Keywords: Subspace clustering, robustness, stability, compressive sensing, sparse

1. Introduction

Subspace clustering is a problem motivated by many real applications. It is now widely known that many high dimensional data including motion trajectories (Costeira and Kanade, 1998), face images (Basri and Jacobs, 2003), network hop counts (Eriksson et al., 2012), movie ratings (Zhang et al., 2012) and social graphs (Jalali et al., 2011) can be modelled as samples drawn from the *union* of multiple low-dimensional linear subspaces (illustrated in Figure 1). Subspace clustering, arguably the most crucial step to understand such data, refers to the task of clustering the data into their original subspaces and uncovers the underlying structure of the data. The partitions correspond to different rigid objects for motion trajectories, different people for face data, subnets for network data, like-minded users in movie database and latent communities for social graph.

Subspace clustering has drawn significant attention in the last decade and a great number of algorithms have been proposed, including Expectation-Maximization-like local optimization algorithms, e.g., K-plane (Bradley and Mangasarian, 2000) and Q-flat (Tseng, 2000), algebraic methods, e.g., Generalized Principal Component Analysis (GPCA) (Vidal et al., 2005), matrix factorization methods (Costeira and Kanade, 1998, 2000), spectral clustering-based methods (Lauer and Schnorr, 2009; Chen and Lerman, 2009), bottom-up local sampling and affinity-based methods (e.g., Yan and Pollefeys, 2006; Rao et al., 2008), and the convex optimization-based methods: namely, Low Rank Representation (LRR) (Liu

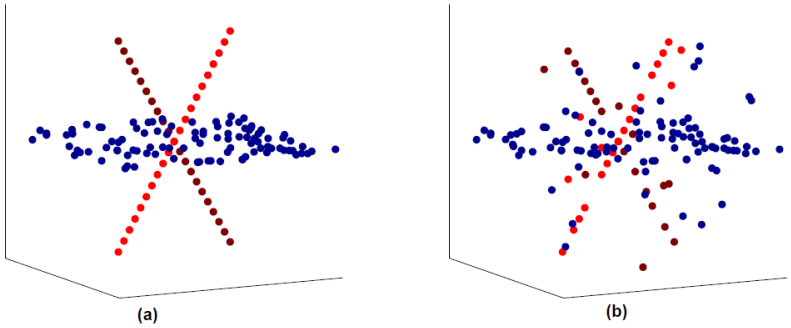


Figure 1: Exact (a) and noisy (b) data in union-of-subspace

et al., 2010, 2013) and Sparse Subspace Clustering (SSC) (Elhamifar and Vidal, 2009, 2013). For a comprehensive survey and comparisons, we refer the readers to the tutorial (Vidal, 2011). Among these algorithms, SSC is known to enjoy superb empirical performance, *even for noisy data*. For example, it is the state-of-the-art algorithm for motion segmentation on Hopkins155 benchmark (Tron and Vidal, 2007; Elhamifar and Vidal, 2009), and has been shown to be more robust than LRR as the number of clusters increase (Elhamifar and Vidal, 2013).

The key idea of SSC is to represent each data point by a sparse linear combination of the remaining data points using ℓ_1 minimization. Without introducing the notations (which is deferred in Section 3), the noiseless and noisy version of SSC solve respectively

$$\min_{c_i} \|c_i\|_1 \quad s.t. \quad x_i = X_{-i}c_i, \qquad \text{and} \qquad \min_{c_i} \|c_i\|_1 + \frac{\lambda}{2}\|x_i - X_{-i}c_i\|^2,$$

for each data column x_i, and the hope is that c_i will be supported only on indices of the data points from the same subspace as x_i. While this formulation is for linear subspaces, affine subspaces can also be dealt with by augmenting data points with an offset variable 1.

Effort has been made to explain the practical success of SSC by analyzing the noiseless version. Elhamifar and Vidal (2010) show that under certain conditions, *disjoint* subspaces (i.e., they are not overlapping) can be exactly recovered. A recent geometric analysis of SSC (Soltanolkotabi and Candes, 2012) broadens the scope of the results significantly to the case when subspaces can be overlapping. However, while these analyses advanced our understanding of SSC, one common drawback is that data points are assumed to be lying *exactly* on the subspaces. This assumption can hardly be satisfied in practice. For example, motion trajectories data are only *approximately* of rank-4 due to perspective distortion of camera, tracking errors and pixel quantization (Costeira and Kanade, 1998); similary, face images are not precisely of rank-9 since human faces are at best *approximated* by a convex body (Basri and Jacobs, 2003).

In this paper, we address this problem and provide a theoretical analysis of SSC with noisy or corrupted data. Our main result shows that a modified version of SSC (see Eq. (3.2)) succeeds when the magnitude of noise does not exceed a threshold determined by a geometric gap between the *inradius* and the *subspace incoherence* (see below for precise definitions). This complements the result of Soltanolkotabi and Candes (2012) that shows

the same geometric gap determines whether SSC succeeds for the noiseless case. Indeed, when the noise vanishes, our results reduce to the noiseless case results of Soltanolkotabi and Candes.

While our analysis is based upon the geometric analysis of Soltanolkotabi and Candes (2012), the analysis is more involved: In SSC, sample points are used as the dictionary for sparse recovery, and therefore noisy SSC requires analyzing a noisy dictionary. We also remark that our results on noisy SSC are *exact*, i.e., as long as the noise magnitude is smaller than a threshold, the recovered subspace clusters are *correct*. This is in sharp contrast to the majority of previous work on structure recovery for noisy data where stability/perturbation bounds are given – i.e., the obtained solution is *approximately* correct, and the approximation gap goes to zero only when the noise diminishes.

Lastly, we remark that an independently developed work (Soltanolkotabi et al., 2014) analyzed the same algorithm *under a statistical model* that generates the data. In contrast, our main results focus on the cases when the data are deterministic. Moreover, when we specialize our general result to the same statistical model, we show that we can handle a significantly larger amount of noise under certain regimes.

The paper is organized as follows. In Section 2, we review previous and ongoing works related to this paper. In Section 3, we formally define the notations, explain our method and the models of our analysis. Then we present our main theoretical results in Section 4 with examples and remarks to explain the practical implications of each theorem. In Section 5 and 6, proofs of the deterministic and randomized results are provided. We then evaluate our method experimentally in Section 7 with both synthetic data and real-life data, which confirms the prediction of the theoretical results. Lastly, Section 8 summarizes the paper and discuss some open problems for future research in the task of subspace clustering.

2. Related works

In this section, we review previous and ongoing theoretical studies on the problem of subspace clustering.

2.1 Nominal performance guarantee for noiseless data

Most previous analyses concern about the nominal performance of a particular subspace clustering algorithm with noiseless data. The focus is to relax the assumptions on the underlying subspaces and data generation.

A number of methods have been shown working under the *independent subspace* assumption including the early factorization-based methods (Costeira and Kanade, 1998; Kanatani, 2001), LRR (Liu et al., 2010) and the initial guarantee of SSC (Elhamifar and Vidal, 2009). Recall that the data points are drawn from a union of subspaces, the *independent subspace* assumption requires each subspace to be linearly independent to the *span* of all other subspaces. Equivalently, this assumption requires the sum of each subspace's dimension to be equal to the dimension of the span of all subspaces. For example, in a two dimensional plane, one can only have 2 independent lines. If there are three lines intersecting at the origin, even if each pair of the lines are independent, they are not considered independent as a whole.

Independent Subspaces	$\dim[\mathcal{S}_1 \otimes ... \otimes \mathcal{S}_L] = \sum_{\ell=1}^{L} \dim[\mathcal{S}_\ell]$.
Disjoint Subspaces	$\mathcal{S}_\ell \cap \mathcal{S}_k = \mathbf{0}$ for all $\{(\ell, k) \vert \ell \neq k\}$.
Overlapping Subspaces	No points lies in $\mathcal{S}_\ell \cap \mathcal{S}_k$ for any $\{(\ell, k) \vert \ell \neq k\}$.

Table 1: Comparison of conditions on the underlying subspaces.

Disjoint subspace assumption only requires pairwise linear independence, and hence is more meaningful in practice. To the best of our knowledge, only GPCA (Vidal et al., 2005) and SSC (Elhamifar and Vidal, 2010, 2013) have been shown to provably handle the data under *disjoint subspace* assumption. GPCA however is not a polynomial time algorithm. Its computational complexity increases exponentially with respect to the number and dimension of the subspaces.

Soltanolkotabi and Candes (2012) developed a geometric analysis that further extends the performance guarantee of SSC, and in particular it covers some cases when the underlying subspaces are slightly *overlapping*, meaning that two subspaces can even share a basis. The analysis reveals that the success of SSC relies on the difference of two geometric quantities (inradius r and incoherence μ) to be greater than 0, which leads to by far the most general and strongest theoretical guarantee for noiseless SSC. A summary of these assumptions on the subspaces and their formal definition are given in Table 1.

We remark that our robust analysis extends from Soltanolkotabi and Candes (2012) and therefore is inherently capable of handling the same range of problems, namely disjoint and overlapping subspaces. This is formalized later in Section 4.

2.2 Robust performance guarantee

Previous studies of the subspace clustering under noise have been mostly empirical. For instance, factorization, spectral clustering and local affinity based approaches, which we mentioned above, are able to produce a (sometimes good) solution even for noisy real data. Convex optimization based approaches like LRR and SSC can be naturally reformulated as a robust method by relaxing the hard equality constraints to a penalty term in the objective function. In fact, the superior results of SSC and LRR on motion segmentation and face clustering data are produced using the robust extension in Elhamifar and Vidal (2009) and Liu et al. (2010) instead of the well-studied noiseless version.

As of writing, there have been very few subspace clustering methods that is guaranteed to work when data are noisy. Besides the conference version of the current paper (Wang and Xu, 2013), an independent work (Soltanolkotabi et al., 2014) also analyzed SSC under noise. Subsequently, there has been noisy guarantees for other algorithms, e.g., thresholding based approach (Heckel and Bölcskei, 2013) and orthogonal matching pursuit (Dyer et al., 2013).

The main difference between our work and (Soltanolkotabi et al., 2014) is that our guarantee works for a more general set of problems when the data and noise may not be random, whereas the key arguments in the proof in Soltanolkotabi et al. (2014) relies on the assumption that data points are uniformly distributed on the unit sphere within each subspace, which corresponds to the "semi-random model" in our paper. As illustrated in Elhamifar and Vidal (2013, Figure 9 and 10), the semi-random model is not a good fit for

	This paper	(Wang and Xu, 2013)	Soltanolkotabi et al. (2014)
Fully deterministic	$O(r(r-\mu))$	$O(r(r-\mu))$	N.A.
Deterministic + random noise	$O((n/d)^{\frac{1}{4}}(r-\mu))$	$O(r-\mu)$	N.A.
Semi-random data + random noise	$O\left(\frac{n^{\frac{1}{4}}}{\sqrt{d}}(1-\frac{\text{aff}}{\sqrt{d}})\right)$	$O\left(\frac{1}{\sqrt{d}}(1-\frac{\text{aff}}{\sqrt{d}})\right)$	$O\left(1-\frac{\text{aff}}{\sqrt{d}}\right)$
Fully-random data + random noise	$O\left(\frac{n^{\frac{1}{4}}}{\sqrt{d}}(1-\frac{\sqrt{d}}{\sqrt{n}})\right)$	$O\left(\frac{1}{\sqrt{d}}(1-\frac{\sqrt{d}}{\sqrt{n}})\right)$	$O\left(1-\frac{\sqrt{d}}{\sqrt{n}}\right)$

Table 2: Comparison of the level of noise tolerable for noisy subspace clustering methods. Note that "aff" is the "unnormalized" affinity defined in (Soltanolkotabi and Candes, 2012)

.

both the motion segmentation and the face clustering datasets, as in these datasets there is a fast decay in the singular values of each subspace. The uniform distribution assumption becomes even harder to justify as the dimension d of each subspace gets larger — a regime where the analysis in (Soltanolkotabi et al., 2014) focuses on.

Moreover, with a minor modification in our analysis that sharpens the bound of the tuning parameter that ensures the solution is non-trivial, we are able to get a result that is stronger than Soltanolkotabi et al. (2014) in cases when the dimension of each subspace $d \leq O(\sqrt{n})$ [1]. This result extends the provably guarantee of SSC to a setting where the signal to noise ratio (SnR) is allowed to go to 0 as the ambient dimension gets large. In summary, we compare our results in terms of the level of noise that can be provably tolerated in Table 2. These comparisons are in the same setting modulo some slight differences in the noise model and successful criteria. It is worth noting that when $d > O(\sqrt{n})$, Soltanolkotabi et al. (2014)'s bound is sharper. We will provide more details in the Appendix.

Lastly, we note that the notion of robustness in this paper is confined to the noise/arbitrary corruptions added to the legitimate data. It is not the robustness against outliers in the data, unless otherwise specified. Handling outliers is a completely different problem. Solutions have been proposed for LRR in Liu et al. (2012) by decomposing a $\ell_{2,1}$ norm column-wise sparse components and for SSC in Soltanolkotabi and Candes (2012) by objective value thresholding. However these results require non-outlier data points to be free of noise, therefore are not comparable to the study in this paper.

3. Problem setup

Notations: We denote the uncorrupted data matrix by $Y \in \mathbb{R}^{n \times N}$, where each column of Y (normalized to unit vector [2]) belongs to a union of L subspaces

$$\mathcal{S}_1 \cup \mathcal{S}_2 \cup ... \cup \mathcal{S}_L.$$

Each subspace $\mathcal{S}_\ell$ is of dimension d_ℓ and contains N_ℓ data samples with $N_1 + N_2 + ... + N_L = N$. We observe the noisy data matrix $X = Y + Z$, where Z is some arbitrary noise

1. Admittedly, (Soltanolkotabi et al., 2014) obtained better noise-tolerance than the comparable result in our conference version (Wang and Xu, 2013).
2. We assume the normalization condition for ease of presentation. Our results can be extened to the case when each column of the noisy data points $X = Y + Z$ is normalized, as well as the case where no normalizing is performed at all, under simple modifications to the conditions.

matrix. Let $Y^{(\ell)} \in \mathbb{R}^{n\times N_\ell}$ denote the selection of columns in Y that belongs to $\mathcal{S}_\ell$, and denote the corresponding columns in X and Z by $X^{(\ell)}$ and $Z^{(\ell)}$ respectively. Without loss of generality, let $X = [X^{(1)}, X^{(2)}, ..., X^{(L)}]$ be ordered. In addition, we use subscript "$-i$" to represent a matrix that excludes column i, e.g., $X^{(\ell)}_{-i} = [x^{(\ell)}_1, ..., x^{(\ell)}_{i-1}, x^{(\ell)}_{i+1}, ..., x^{(\ell)}_{N_\ell}]$. Calligraphic letters such as $\mathcal{X}, \mathcal{Y}_\ell$ represent the set containing all columns of the corresponding matrix (e.g., X and $Y^{(\ell)}$).

For any matrix X, $\mathcal{P}(X)$ represents the symmetrized convex hull of its columns, i.e., $\mathcal{P}(X) = \text{conv}(\pm\mathcal{X})$. Also let $\mathcal{P}^{(\ell)}_{-i} := \mathcal{P}(X^{(\ell)}_{-i})$ and $\mathcal{Q}^{(\ell)}_{-i} := \mathcal{P}(Y^{(\ell)}_{-i})$ for short. $\mathbb{P}_{\mathcal{S}}$ and $\text{Proj}_{\mathcal{S}}$ denote respectively the projection matrix and projection operator (acting on a set) to subspace $\mathcal{S}$. Throughout the paper, $\|\cdot\|$ represents 2-norm for vectors and operator norm for matrices; other norms will be explicitly specified (e.g., $\|\cdot\|_1, \|\cdot\|_\infty$).

Method: Original SSC solves the linear program

$$\min_{c_i} \|c_i\|_1 \quad s.t. \quad x_i = X_{-i}c_i, \tag{3.1}$$

for each data point x_i. Solutions are arranged into matrix $C = [c_1, ..., c_N]$, then spectral clustering techniques such as Ng et al. (2002) are applied on the affinity matrix $W = |C| + |C|^T$ ($|\cdot|$ represents entrywise absolute value). Note that when $Z \neq 0$, this method breaks down: indeed (3.1) may even be infeasible.

To handle noisy X, a natural extension is to relax the equality constraint in (3.1) and solve the following unconstrained minimization problem instead (Elhamifar and Vidal, 2013):

$$\min_{c_i} \|c_i\|_1 + \frac{\lambda}{2}\|x_i - X_{-i}c_i\|^2. \tag{3.2}$$

We will focus on Formulation (3.2) in this paper. Notice that (3.2) coincides with standard LASSO. Yet, since our task is subspace clustering, the analysis of LASSO (mainly for the task of support recovery) does not extend to SSC. In particular, existing literature for LASSO to succeed requires the dictionary X_{-i} to satisfy the Restricted Isometry Property (RIP for short; Candès, 2008) or the Null-space property (Donoho et al., 2006), but neither of them is satisfied in the subspace clustering setup.[3]

In the subspace clustering task, there is no single "ground-truth" C to compare the solution against. Instead, the algorithm succeeds if each sample is expressed as a linear combination of samples belonging to the same subspace, as the following definition states.

Definition 1 (LASSO Subspace Detection Property) *We say the subspaces $\{\mathcal{S}_\ell\}_{\ell=1}^k$ and noisy sample points X from these subspaces obey LASSO subspace detection property with parameter λ, if and only if it holds that for all i, the optimal solution c_i to* (3.2) *with parameter λ satisfies:*

(1) c_i is not a zero vector, i.e., the solution is non-trivial, (2) Nonzero entries of c_i correspond to only columns of X sampled from the same subspace as x_i.

This property ensures that the output matrix C and (naturally) the affinity matrix W are exactly block diagonal with each subspace cluster represented by a disjoint block. The

3. As a simple illustrative example, suppose there exists two identical columns in X_{-i}, which violates RIP for 2-sparse signal and has maximum incoherence $\mu(X_{-i}) = 1$.

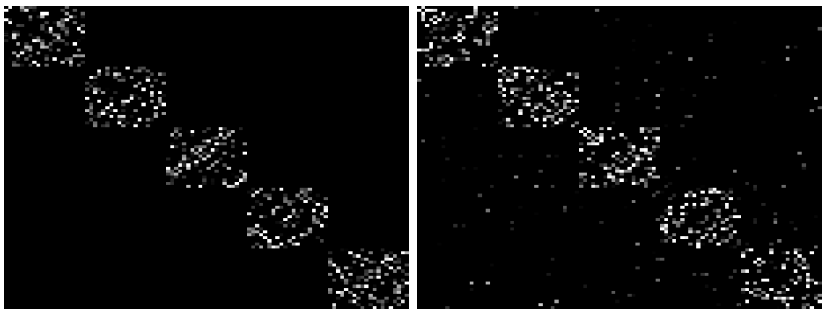


Figure 2: Illustration of LASSO-Subspace Detection Property/Self-Expressiveness Property. **Left:** SEP holds. **Right:** SEP is violated even though spectral clustering is likely to cluster this affinity graph perfectly into 5 blocks.

property is illustrated in Figure 2. For convenience, we will refer to the second requirement alone as "*Self-Expressiveness Property*" (SEP), as defined in Elhamifar and Vidal (2013).

Note that the LASSO Subspace Detection Property is a strong condition. In practice, spectral clustering does not require the exact block diagonal structure for perfect segmentation (check Figure 9 in our simulation section for details). A caveat is that it is also not sufficient for perfect segmentation, since it does not guarantee each diagonal block forms a connected component. This is a known problem for SSC (Nasihatkon and Hartley, 2011), although we observe that in practice graph connectivity is usually not a big issue. Proving the high-confidence connectivity (even under probabilistic models) remains an open problem, except for the almost trivial cases when the subspaces are independent (Liu et al., 2013; Wang et al., 2013).

Models of analysis: Our objective here is to provide sufficient conditions upon which the LASSO subspace detection properties hold in the following four models. Precise definition of the noise models will be given in Section 4.

- fully deterministic model;
- deterministic data with random noise;
- semi-random data with random noise;
- fully random model.

4. Main results

4.1 Deterministic model

We start by defining two concepts adapted from the original proposal of Soltanolkotabi and Candes (2012).

Definition 2 (Projected Dual Direction) *Let ν be the optimal solution to the dual optimization program*[4]

$$\max_{\nu} \langle x, \nu \rangle - \frac{1}{2\lambda}\nu^T\nu, \quad \textit{subject to:} \quad \|X^T\nu\|_\infty \leq 1;$$

4. This definition is related to (5.3), the dual problem of (3.2), which we will define in the proof.

and $\mathcal{S}$ is a low-dimensional subspace. The projected dual direction *v is defined as*

$$v(x, X, \mathcal{S}, \lambda) \triangleq \frac{\mathbb{P}_{\mathcal{S}}\nu}{\|\mathbb{P}_{\mathcal{S}}\nu\|}.$$

Definition 3 (Projected Subspace Incoherence Property) *Compactly denote projected dual direction $v_i^{(\ell)} = v(x_i^{(\ell)}, X_{-i}^{(\ell)}, \mathcal{S}_\ell, \lambda)$ and $V^{(\ell)} = [v_1^{(\ell)}, ..., v_{N_\ell}^{(\ell)}]$. We say that vector set $\mathcal{X}_\ell$ is μ-incoherent to other points if*

$$\mu \geq \mu(\mathcal{X}_\ell) := \max_{y \in \mathcal{Y} \setminus \mathcal{Y}_\ell} \|{V^{(\ell)}}^T y\|_\infty.$$

Here, μ measures the incoherence between corrupted subspace samples $\mathcal{X}_\ell$ and clean data points in other subspaces (illustrated in Figure 4). As $\|y\| = 1$ by the normalization assumption, the range of μ is $[0, 1]$. In case of random subspaces in high dimension, μ is close to zero. Moreover, as we will see later, for deterministic subspaces and random data points, μ is proportional to their expected angular distance (measured by cosine of canonical angles).

Definition 2 and 3 differ from the *dual direction* and *subspace incoherence property* of Soltanolkotabi and Candes (2012) in that we require a projection to a particular subspace to cater to the analysis of the noise case. Also, since they reduce to the original definitions when data are noiseless and $\lambda \to \infty$, these definitions can be considered as a generalization of their original version.

Definition 4 (inradius) *The inradius of a convex body $\mathcal{P}$, denoted by $r(\mathcal{P})$, is defined as the radius of the largest Euclidean ball inscribed in $\mathcal{P}$.*

The inradius of a $\mathcal{Q}_{-i}^{(\ell)}$ describes the dispersion of the data points. Well-dispersed data lead to larger inradius and skewed/concentrated distribution of data have small inradius. An illustration is given in Figure 3.

Definition 5 (Deterministic noise model) *Consider arbitrary additive noise Z to Y, each column z_i is bounded by the two quantities below:*

$$\delta := \max_i \|z_i\|, \qquad \delta_1 := \max_{i,\ell} \|\mathbb{P}_{\mathcal{S}_\ell} z_i\|,$$

As we assume the uncorrupted data point y has unit norm, δ essentially describes the amount of allowable relative error.

Theorem 6 *Under the deterministic noise model, compactly denote*

$$\mu_\ell := \mu(\mathcal{X}_\ell), \qquad r_\ell := \min_{\{i: x_i \in \mathcal{X}_\ell\}} r(\mathcal{Q}_{-i}^{(\ell)}), \qquad r := \min_{\ell=1,...,L} r_\ell.$$

If $\mu_\ell < r_\ell$ for each $\ell = 1, ..., L$, furthermore

$$\delta \leq \min_{\ell=1,...,L} \frac{r(r_\ell - \mu_\ell)}{2 + 7r_\ell}$$

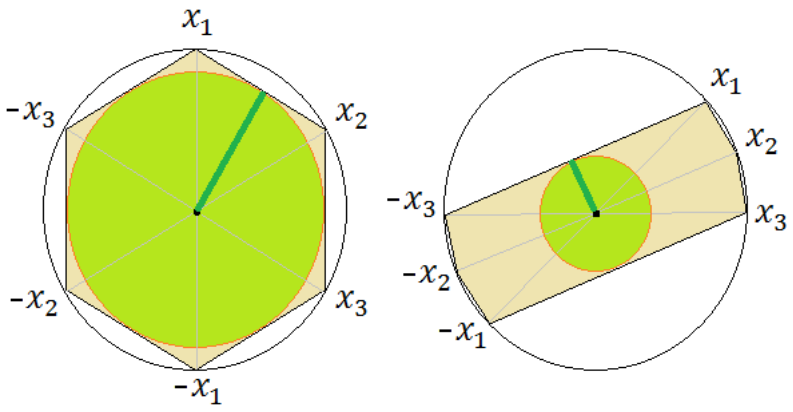


Figure 3: Illustration of inradius and data distribution. The inradius measures how well data points represent a subspace.

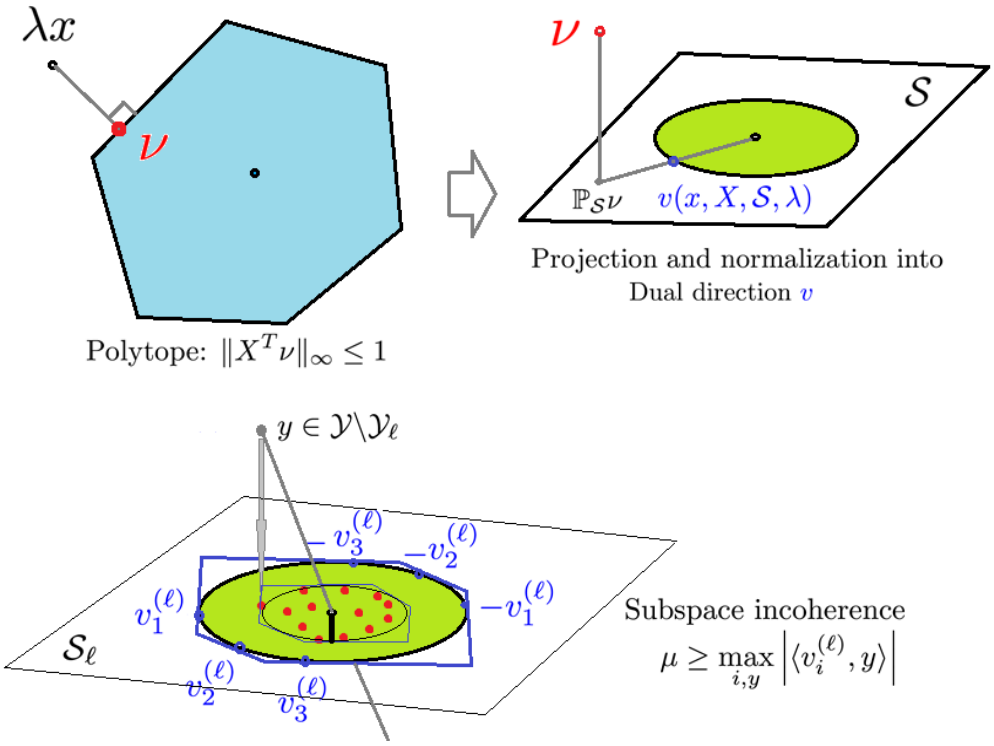


Figure 4: Illustrations of the projected dual direction and subspace incoherence property. The projected dual direction in Definition 2 is essentially an Euclidean projection to the polytope, followed by a projection to the subspace and normalization. There is a dual direction associated with each data point in the subspace. Jointly, $\left\{x \middle| \max_i \left|\langle v_i^{(\ell)}, x\rangle\right| \le \mu\right\}$ defines a polygon in the subspace $\mathcal{S}_\ell$, and subspace incoherence μ is given by the smallest such polytope that contains the projections of all external point y into the this subspace.

then LASSO subspace detection property holds for all weighting parameter λ in the range

$$\frac{1}{r-2\delta-\delta^2} < \lambda < \min_{\ell=1,..,L} \left\{ \frac{r_\ell - \mu_\ell - 2\delta_1}{\delta(1+\delta)(2+r_\ell-\delta_1)} \right\}$$

which is guaranteed to be non-empty.

We now offer some discussions of the theorem and the proof will be given in Section 5.

Noiseless case. When $\delta = 0$, i.e., there is no noise, the condition reduces to $\mu_\ell < r_\ell$, which coincides with the result in Soltanolkotabi and Candes (2012). The exact LP formulation (3.1) is equivalent to $\lambda \to \infty$. Our result implies that unconstrained LASSO formulation (3.2) works for any $\lambda > \frac{1}{r}$.

Signal-to-Noise Ratio. Condition $\delta \leq \frac{r(r-\mu)}{2+7r}$ can be interpreted as the breaking point under increasing magnitude of attack. This suggests that SSC by (3.2) is provably robust to arbitrary noise having signal-to-noise ratio (SNR) greater than $\Theta\big(\frac{1}{r(r-\mu)}\big)$. (Notice that $0 < r < 1$, and hence $7r + 2 = \Theta(1)$.)

Tuning parameter λ. The range of the parameter λ in the theorem depends on unknown parameters μ, r and δ, and therefore cannot be used in practice to choose the parameter in practice. It does however justify that when δ is small, the range of λ that Lasso-SSC works is large, therefore not hard to tune. In practice, we do not need to know λ in prior. One approach is to trace the Lasso path (Tibshirani et al., 2013) until we have about k non-zero entries in the coefficient vector. If we would like to use a single λ for all columns, a good point to start is to take λ to be in the order of $O\big(\frac{1}{\min_j \max_{i\neq j} |x_i^T x_j|}\big)$, this ensures the solution to be at least non-trivial.

Agnostic subspace clustering. The robustness to deterministic error is important, since in practice the union-of-subspace structures are usually only good approximations. If each subspace has decaying singular values (e.g., motion segmentation, face clustering (Elhamifar and Vidal, 2013) and hybrid system identification(Vidal et al., 2003)), the deterministic guarantee allows for the flexibility in choosing the cut-off points, e.g., take 90% of the energy as signal and treat the remaining spectrum as noise. If one keeps a smaller number of singular values (a smaller subspace dimension), the inradius will likely to be larger [5], although the noise level also increases. It is possible that the conditions in Theorem 6 are satisfied for some decomposition (e.g., those with a large spectral gap) but not others. The nice thing is that this is not a tuning parameter, but rather a theoretical property that remains agnostic to the users. In fact, the algorithm will be provably effective as long as the conditions are satisfied for any signal noise decomposition (not restricted to rank-projection). None of these is possible if distributional assumptions are made to either the data or the noise.

4.2 Randomized models

We further analyze three randomized models with increasing level of randomness.

5. A formal relationship between inradius and smallest singular value is described in (Wang et al., 2013).

- **Determinitic+Random Noise.** Subspaces and samples in subspace are arbitrary; the noise obeys the Random Noise model (Definition 7).

- **Semi-random+Random Noise.** Subspace is deterministic, but samples in each subspace are drawn iid uniformly from the intersection of the unit sphere and the subspace; the noise obeys the Random Noise model.

- **Fully random.** Both subspace and samples are drawn uniformly at random from their respective domains; the noise is iid Gaussian.

In each of these models, we improve the performance guarantee over our conference version (Wang and Xu, 2013). In the most well-studied semi-random model, we are able to handle cases where the noise level is much larger than the signal, and hence improves upon the best known result for SSC Soltanolkotabi et al. (2014). A detailed comparison of the noise tolerance of these methods is given in Table 2.

Definition 7 (Random noise model) *Our random noise model is defined to be any additive Z that is (1) columnwise iid; (2) spherical symmetric; and (3) $\|z_i\| \leq \delta$ for all $i = 1, ..., N$ with probability at least $1 - 1/N$.*

A good example of our random noise model is iid Gaussian noise. Let each entry $Z_{ij} \sim N(0, \sigma^2/n)$. It is known that (see Lemma 18) for some constant C

$$\mathbb{P}\left(\delta := \max_i \|z_i\| > \sqrt{1 + \frac{6\log N}{n}}\sigma\right) \leq C/N^2.$$

Theorem 8 (Deterministic+Random Noise) *Under random noise model, compactly denote r_ℓ, r and μ_ℓ as in Theorem 6, furthermore let*

$$\epsilon := \sqrt{\frac{6\log N}{n - \max_\ell d_\ell}} \leq \sqrt{\frac{C\log(N)}{n}}.$$

If $\mu_\ell < r_\ell$ for all $\ell = 1, ..., k$,

$$\epsilon\delta < \min_{\ell=1,...,L} \frac{r_\ell - \mu_\ell}{2\sqrt{d_\ell} + 2}, \qquad \text{and} \qquad \epsilon\delta(1+\delta) < \min_{\ell=1,...,L} \frac{r(r_\ell - \mu_\ell)}{4r_\ell + 6},$$

then with probability at least $1 - 9/N$, LASSO subspace detection property holds for all weighting parameter λ in the range

$$\frac{1}{r - 2\epsilon\delta - \epsilon\delta^2} < \lambda < \min_{\ell=1,...,L} \left\{ \frac{r_\ell - \mu_\ell - \delta\epsilon - \delta\sqrt{d_\ell}\epsilon}{\epsilon\delta(1+\delta)(3 + r_\ell - \delta\sqrt{d_\ell}\epsilon)} \right\} \tag{4.1}$$

which is guaranteed to be non-empty.

Low SnR paradigm. Compared to Theorem 6, Theorem 8 considers a more benign noise which leads to a stronger result. In particular, without assuming any statistical model on how data are generated, we show that Lasso-SSC is able to tolerate noise of level $O\left((\frac{n}{\log N})^{1/4}(r(r_\ell-\mu_\ell))^{1/2}\right)$ or $O\left((\frac{n}{d\log N})^{1/2}(r_\ell-\mu_\ell)\right)$ (whichever is smaller). This extends SSC's guarantee with deterministic data to cases where the noise can be significantly larger than the signal. In fact, the SnR can go to 0 as the ambient dimension gets large.

On the other hand, Theorem 8 shows that Lasso-SSC is able to tolerate a constant level of noise when the geometric gap $r_\ell - \mu_\ell$ is as small as $O(\sqrt{d/n})$. This is arguably near-optimal (when d is small) as the projection of a constant-level random noise into a d-dimensional subspace has an expected magnitude of the same order, which could easily close up the small geometric gap for some non-trivial probability if the noise is much larger.

Margin of error. Since the bound depends critically on $(r_\ell - \mu_\ell)$ – the difference of inradius and incoherence – which is the geometric gap that appears in the noiseless guarantee of Soltanolkotabi and Candes (2012). We will henceforth call this gap the *margin of error*.

We now analyze this margin of error under different generative models. We start from the semi-random model, where the distance between two subspaces is measured as follows.

Definition 9 *The* affinity *between two subspaces is defined by:*

$$\text{aff}(\mathcal{S}_k,\mathcal{S}_\ell)=\sqrt{\cos^2\theta_{k\ell}^{(1)}+...+\cos^2\theta_{k\ell}^{(\min(d_k,d_\ell))}},$$

where $\theta_{k\ell}^{(i)}$ is the i^{th} canonical angle between the two subspaces. Let U_k and U_ℓ be a set of orthonormal bases of each subspace, then $\text{aff}(\mathcal{S}_k,\mathcal{S}_\ell)=\|U_k^TU_\ell\|_F$.

When data points are randomly sampled from each subspace, the geometric entity $\mu(\mathcal{X}_\ell)$ can be expressed using this (more intuitive) subspace affinity, which leads to the following theorem.

Theorem 10 (Semi-random model+random noise) *Under the semi-random model with random noise, there exists a non-empty range of λ such that LASSO subspace detection property holds with probability* $1-\frac{9}{N}-\frac{1}{L^2}\sum_{\ell\neq\ell'}\frac{1}{(N_\ell+1)N_{\ell'}}e^{-\frac{t}{4}}-6\sum_{\ell=1}^{L}(e^{\gamma_1(n-d_\ell)}+e^{\gamma_2 d_\ell}+e^{-\sqrt{N_\ell d_\ell}})$ *as long as the noise level obeys*

$$\delta(1+\delta)\leq\max_{\ell,\ell'}\sqrt{\frac{n-d}{6\log N}}\frac{\sqrt{\log\kappa}}{40K_2\sqrt{dd_\ell}}\left(1-\frac{K_1K_2\text{aff}(\mathcal{S}_\ell,\mathcal{S}_{\ell'})}{\sqrt{d_{\ell'}}}\right),$$

where $K_1:=(t\log[(N_\ell+1)N_{\ell'}]+\log L)$, $K_2:=4\sqrt{\frac{1}{\log\kappa_\ell}}$, $\kappa_\ell:=N_\ell/d_\ell$, $\frac{\log\kappa}{d}:=\min_\ell\frac{\log\kappa_\ell}{d_\ell}$, and γ_1,γ_2 are absolute constants.

The proof is essentially substituting the incoherence and inradius parameters in Theorem 8 with meaningful bounds, so Thereom 10 can be regarded as a corollary of Theorem 8.

Overlapping subspaces. Similar to the results in Soltanolkotabi and Candes (2012), Theorem 10 demonstrates that LASSO-SSC can handle overlapping subspaces with noisy samples. By Definition 9, $\text{aff}(\mathcal{S}_k,\mathcal{S}_\ell)$ can be small even if $\mathcal{S}_k$ and $\mathcal{S}_\ell$ share a basis.

Application	Cluster rank
3D motion segmentation (Costeira and Kanade, 1998)	rank = 4
Face clustering (with shadow) (Basri and Jacobs, 2003)	rank = 9
Diffuse photometric face (Zhou et al., 2007)	rank = 3
Network topology discovery (Eriksson et al., 2012)	rank = 2
Hand writing digits (Hastie and Simard, 1998)	rank = 12
Social graph clustering (Jalali et al., 2011)	rank = 1

Table 3: Rank of real subspace clustering problems

Comparison to Soltanolkotabi et al. (2014). In the high dimensional setting when $n \gg d$, our result is able to handle the low SnR regime when $\delta = \Theta(n^{1/4}/d^{1/2})$, while Soltanolkotabi et al. (2014) needs δ to be bounded by a small constant.

In the case when d is a constant fraction of n, however, our bound is worse by a factor of $\sqrt{d}$. Soltanolkotabi et al. (2014) is still able to handle a small constant noise while we needs $\delta < O(\frac{1}{\sqrt{d}})$. The suboptimal bound might be due to the fact that we are simply developing the theorem for the semirandom model as a corollary of Theorem 8 and haven not fully exploit the structure of the semi-random model in the proof.

We now turn to the fully random case.

Theorem 11 (Fully random model) *Suppose there are L subspaces each with dimension d, chosen independently and uniformly at random. For each subspace, $\kappa d + 1$ points are chosen independently and uniformly from the unit sphere inside each subspace. Each measurement is corrupted by iid Gaussian noise $\sim N(0, \sigma^2/n)$. Furthermore, if*

$$d < \frac{c(\kappa)^2 \log \kappa}{24 \log N} n, \qquad \text{and} \qquad \sigma(1+\sigma) < \frac{c(\kappa)^2 \log \kappa}{20} \frac{\sqrt{n}}{d},$$

then with probability at least $1 - \frac{10}{N} - Ne^{-\sqrt{\kappa} d}$, the LASSO subspace detection property holds for any λ in the range

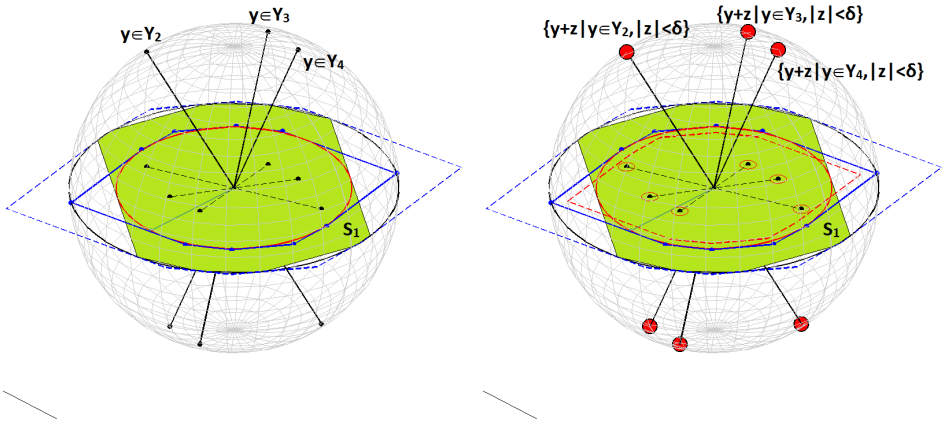
$$\frac{C_1 \sqrt{d}}{c(\kappa)\sqrt{\log \kappa}} < \lambda < \frac{C_2 c(\kappa) \sqrt{n \log \kappa}}{\sigma \sqrt{d \log N}}, \tag{4.2}$$

which is guaranteed to be non-empty. Here, C_1, C_2 are absolute constants.

The results under this simple model are very interpretable. It provides intuitive guideline in how robustness of Lasso-SSC change with respect to the various parameters of the data. One one hand, it is sensitive to the dimension of each subspace d, since the $\sigma \leq \tilde{\Theta}(\frac{n^{1/4}}{\sqrt{d}})$. This dependence on subspace dimension d is not a critical limitation as most interesting applications indeed have very low subspace-dimension, as summarized in Table 3. On the other hand, the dependence on the number of subspaces L (in both $\log \kappa$ and $\log N$ since $N = L(\kappa d + 1)$) is only logarithmic. This suggests that SSC is robust even when there are many clusters, and $Ld \gg n$.

4.3 Geometric interpretations

A geometric illustration of the condition in Theorem 6 is given in Figure 5 in comparison to the geometric separation condition in the noiseless case.



(a) Noiseless SSC (b) Noisy LASSO-SSC

Figure 5: Geometric interpretation and comparison of the noiseless SSC (**Left**) and noisy LASSO-SSC (**Right**).

The left pane depicts the separation condition $\mu_\ell \leq r_\ell$ in Theorem 2.5 of Soltanolkotabi and Candes (2012). The blue polygon represents the the intersection of halfspaces defined with dual directions that are also the tangent to the red inscribing sphere. More precisely, this is $\left\{x \in \mathcal{S}_\ell \middle| |\langle v_i^\ell, x\rangle| \leq r_\ell\right\}$. From our illustration of μ in Figure 4, we can easily tell that $\mu_\ell \leq r_\ell$ if and only if the projection of external data points fall inside this solid blue polygon. We call this solid blue polygon the successful region.

The right pane illustrates our guarantee of Theorem 6 under bounded deterministic noise. The successful condition requires that the whole red ball (analogous to uncertainty set in Robust Optimization (Ben-Tal and Nemirovski, 1998; Bertsimas and Sim, 2004)) around each external data point to fall inside the dashed red polygon, which is smaller than the blue polygon by a factor related to the noise level and the inradius.

The successful region is affected by the noise because the design matrix is also arbitrarily perturbed and the dual solution is no longer within each subspace $\mathcal{S}_\ell$. Specifically, as will become clear in the proof, the key of showing SEP boils down to proving $\langle \nu_i^{(\ell)}, x_j\rangle < 1$ for all pairs of $(\nu_i^{(\ell)}, x_j)$ where

$$\nu_i^{(\ell)} = \arg\max_{\nu} \langle \nu, x_i^{(\ell)}\rangle - \frac{1}{2\lambda}\|\nu\|^2 \text{ s.t. } \|\nu^T X_{-i}^{(\ell)}\|_\infty \leq 1,$$

and x_j is any point from another subspace. In the noiseless case we can always take $\nu_i^{(\ell)} \in \mathcal{S}_\ell$ and $\langle \nu_i^{(\ell)}, x_j\rangle \leq \frac{\mu_\ell}{r_\ell}$. For noisy data and Lasso-SSC, we can no longer do that. In fact, for any fixed λ, the dual solution will be uniquely determined by a projection of $\lambda x_i^{(\ell)}$ on to the feasible region $\|\nu^T X_{-i}^{(\ell)}\|_\infty \leq 1$ (see the first pane of Figure 4). The absolute value of the inner product $\langle \nu_i^{(\ell)}, x_j\rangle$ will depend on the magnitude of the dual solution, especially its component perpendicular to the current subspace. Indeed by carefully choosing the error, we can make $\mathbb{P}_{\mathcal{S}_\ell^\perp}\nu$ very correlated with some external data point x_j.

To illustrate this further, we plot the shape of this feasible region in 3D (see Figure 6(b)). From the feasible region alone, it seems that the magnitude of dual variable can potentially be quite large. Luckily, the quadratic penalty in the objective function allows us to exploit the optimality of the solution ν and bound the "out-of-subspace" component of ν, which results in a much smaller region where the solution can potentially be (given in Figure 6(c)). The region for the "in-subspace" component is also smaller as is shown in Figure 7. A more detailed argument of this is given in Section 5.3 of the proof.

Admittedly, the geometric interpretation under noise is slightly messier than the noiseless case, but it is clear that the largest deterministic noise Lasso-SSC can tolerate must be smaller than geometric gap $r_\ell - \mu_\ell$. Theorem 6 show that a sufficient condition is $\delta \leq O(r(r_\ell - \mu_\ell))$. It remains unclear whether this gap can be closed without additional assumptions.

Finally, we note that for the random noise model in Theorem 8, the geometric interpretation is similar, except that the impact of the noise is weakened. Thanks to the randomness and the corresponding concentration of measure, we may bound the reduction of the successful region with a much smaller value comparing to the adversarial noise case.

5. Proof of the Deterministic Result

In this section, we provide the proof for Theorem 6.

Instead of analyzing (3.2) directly, we consider an equivalent constrained version by introducing slack variable e_i:

$$\mathbf{P}_0: \quad \min_{c_i, e_i} \|c_i\|_1 + \frac{\lambda}{2}\|e_i\|^2 \quad s.t. \quad x_i^{(\ell)} = X_{-i}c_i + e_i. \tag{5.1}$$

The constraint can be rewritten as

$$y_i^{(\ell)} + z_i^{(\ell)} = (Y_{-i} + Z_{-i})c_i + e_i. \tag{5.2}$$

The dual program of (5.1) is:

$$\mathbf{D}_0: \quad \max_{\nu} \langle x_i, \nu \rangle - \frac{1}{2\lambda}\nu^T\nu \quad s.t. \quad \|(X_{-i})^T\nu\|_\infty \leq 1. \tag{5.3}$$

Recall that we want to establish the conditions on noise magnitude δ, structure of the data (μ and r in the deterministic model and affinity in the semi-random model), and ranges of valid λ such that by Definition 1, the solution c_i is *non-trivial* and has support indices inside the column set $X_{-i}^{(\ell)}$ (i.e., satisfies SEP).

The proof is hence organized into three main steps:

(1) Proving SEP by duality. First we establish a set of conditions on the optimal dual variable of D_0 corresponding to all primal solutions satisfying SEP. Then we construct such a dual variable ν as a certificate of proof. This is presented in Section 5.1, 5.2 and 5.3.

(2) Proving non-trivialness by showing that the optimal value is smaller than the value of the trivial solution (i.e., $c^* = 0$ and $e^* = x_i^{(\ell)}$). This step is given in Section 5.4.

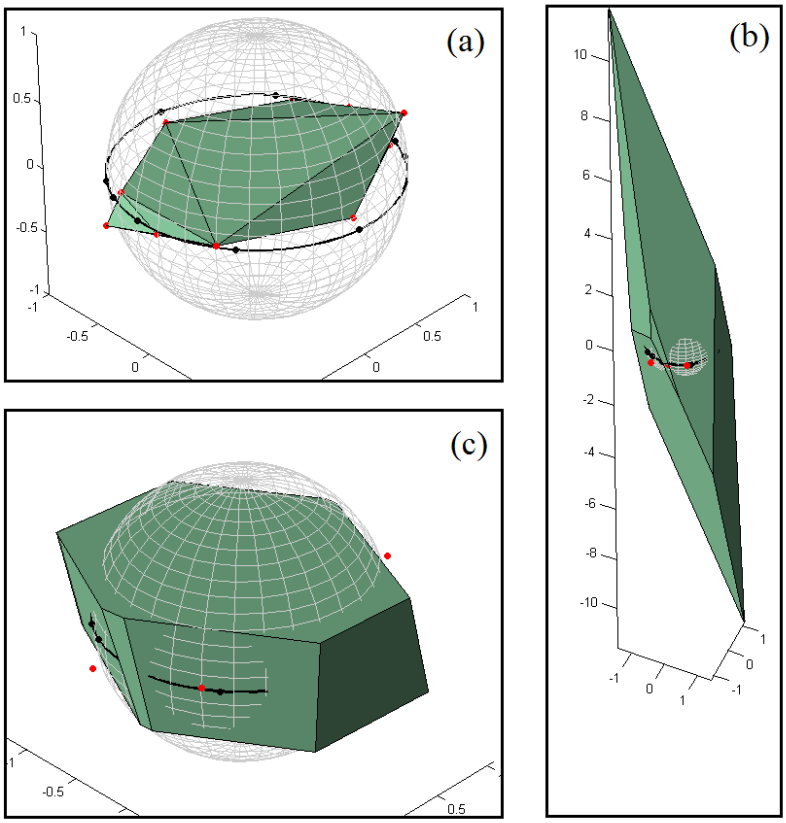


Figure 6: Illustration of **(a)** the convex hull of noisy data points, **(b)** its polar set and **(c)** the intersection of polar set and $\|\nu_2\|$ bound. The polar set (b) defines the feasible region of (5.7). It is clear that ν_2 can take very large value in (b) if we only consider feasibility. By considering optimality, we know the optimal ν must be inside the region in (c).

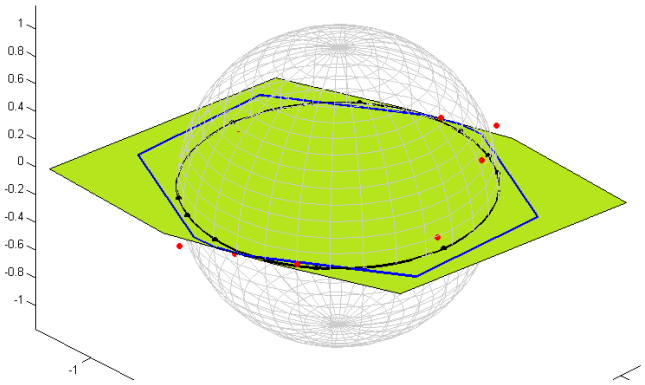


Figure 7: The projection of the polar set (the green area) in comparison to the projection of the polar set with $\|\nu_2\|$ bound (the blue polygon). It is clear that the latter is much smaller.

(3) Showing the existence of a proper λ. As it will be made clear later, conditions for (1) include $\lambda < A$ and (2) requires $\lambda > B$ for some expression A and B. Then it is natural to request $B < A$, so that a valid λ exists. It turns out that this condition boils down to $\delta < C$ for some expression C. This argument is carried over in Section 5.5.

5.1 Optimality Condition

Consider a general convex optimization problem:

$$\min_{c,e} \|c\|_1 + \frac{\lambda}{2}\|e\|^2 \qquad s.t. \quad x = Ac + e. \tag{5.4}$$

We state Lemma 12, which extends Lemma 7.1 in Soltanolkotabi and Candes (2012).

Lemma 12 *Consider a vector $y \in \mathbb{R}^d$ and a matrix $A \in \mathbb{R}^{d\times N}$. If there exists a triplet (c, e, ν) obeying $y = Ac + e$ and c has support $S \subseteq T$, furthermore the dual certificate vector ν satisfies*

$$A_s^T\nu = sgn(c_S), \quad \nu = \lambda e, \quad \|A_{T\cap S^c}^T\nu\|_\infty \le 1, \quad \|A_{T^c}^T\nu\|_\infty < 1,$$

then any optimal solution (c^*, e^*) to (5.4) obeys $c^*_{T^c} = 0$.

Proof For optimal solution (c^*, e^*), we have:

$$\begin{aligned}
&\|c^*\|_1 + \frac{\lambda}{2}\|e^*\|^2 \\
=&\|c^*_S\|_1 + \|c^*_{T\cap S^c}\|_1 + \|c^*_{T^c}\|_1 + \frac{\lambda}{2}\|e^*\|^2 \\
\ge&\|c_S\|_1 + \langle sgn(c_S), c^*_S - c_S\rangle + \|c^*_{T\cap S^c}\|_1 + \|c^*_{T^c}\|_1 + \frac{\lambda}{2}\|e\|^2 + \langle \lambda e, e^* - e\rangle \\
=&\|c_S\|_1 + \langle \nu, A_S(c^*_S - c_S)\rangle + \|c^*_{T\cap S^c}\|_1 + \|c^*_{T^c}\|_1 + \frac{\lambda}{2}\|e\|^2 + \langle \nu, e^* - e\rangle \\
=&\|c_S\|_1 + \frac{\lambda}{2}\|e\|^2 + \|c^*_{T\cap S^c}\|_1 - \langle \nu, A_{T\cap S^c}(c^*_{T\cap S^c})\rangle + \|c^*_{T^c}\|_1 - \langle \nu, A_{T^c}(c^*_{T^c})\rangle.
\end{aligned} \tag{5.5}$$

To see $\frac{\lambda}{2}\|e^*\|^2 \ge \frac{\lambda}{2}\|e\|^2 + \langle \lambda e, e^* - e\rangle$, note that the right hand side equals to $\lambda\left(-\frac{1}{2}e^Te + (e^*)^Te\right)$, which takes a maximal value of $\frac{\lambda}{2}\|e^*\|^2$ when $e = e^*$. The last equation holds because both (c, e) and (c^*, e^*) are feasible solution, such that $\langle \nu, A(c^* - c)\rangle + \langle \nu, e^* - e\rangle = \langle \nu, Ac^* + e^* - (Ac + e)\rangle = 0$. Also, note that $\|c_S\|_1 + \frac{\lambda}{2}\|e\|^2 = \|c\|_1 + \frac{\lambda}{2}\|e\|^2$.

With the inequality constraints of ν given in the lemma statement, we have

$$\langle \nu, A_{T\cap S^c}(c^*_{T\cap S^c})\rangle = \langle A^T_{T\cap S^c}\nu, (c^*_{T\cap S^c})\rangle \le \|A^T_{T\cap S^c}\nu\|_\infty \|c^*_{T\cap S^c}\|_1 \le \|c^*_{T\cap S^c}\|_1.$$

Substitute into (5.5), we get:

$$\|c^*\|_1 + \frac{\lambda}{2}\|e^*\|^2 \ge \|c\|_1 + \frac{\lambda}{2}\|e\|^2 + (1 - \|A^T_{T^c}\nu\|_\infty)\|c^*_{T^c}\|_1,$$

where $(1 - \|A^T_{T^c}\nu\|_\infty)$ is strictly greater than 0.

Using the fact that (c^*, e^*) is an optimal solution, $\|c^*\|_1 + \frac{\lambda}{2}\|e^*\|^2 \leq \|c\|_1 + \frac{\lambda}{2}\|e\|^2$. Therefore, $\|c^*_{T^c}\|_1 = 0$ and (c, e) is also an optimal solution. This concludes the proof. ■

The next step is to apply Lemma 12 with $x = x_i^{(\ell)}$ and $A = X_{-i}$ and then construct a triplet (c, e, ν) such that dual certificate ν satisfying all conditions and c satisfies SEP. Then we can conclude that all optimal solutions of (5.1) satisfy SEP.

5.2 Construction of Dual Certificate

To construct the dual certificate, we consider the following *fictitious* optimization problem (and its dual) that explicitly requires that all feasible solutions satisfy SEP[6] (note that one can not solve such problem in practice without knowing the subspace clusters, and hence the name "fictitious").

$$\mathbf{P}_1: \quad \min_{c_i^{(\ell)}, e_i} \|c_i^{(\ell)}\|_1 + \frac{\lambda}{2}\|e_i\|^2 \quad s.t. \quad y_i^{(\ell)} + z_i = (Y_{-i}^{(\ell)} + Z_{-i}^{(\ell)})c_i^{(\ell)} + e_i; \tag{5.6}$$

$$\mathbf{D}_1: \quad \max_{\nu} \langle x_i^{(\ell)}, \nu\rangle - \frac{1}{2\lambda}\nu^T\nu \quad s.t. \quad \|(X_{-i}^{(\ell)})^T\nu\|_\infty \leq 1. \tag{5.7}$$

This optimization problem is feasible because $y_i^{(\ell)} \in span(Y_{-i}^{(\ell)}) = \mathcal{S}_\ell$ so any $c_i^{(\ell)}$ obeying $y_i^{(\ell)} = Y_{-i}^{(\ell)}c_i^{(\ell)}$ and corresponding $e_i = z_i - Z_{-i}^{(\ell)}c_i^{(\ell)}$ is a pair of feasible solution. Then by strong duality, the dual program is also feasible, which implies that for every optimal solution (c, e) of (5.6) with c supported on S, there exist ν satisfying:

$$\left\{\begin{array}{c} \|((Y_{-i}^{(\ell)})_{S^c}^T + (Z_{-i}^{(\ell)})_{S^c}^T)\nu\|_\infty \leq 1, \quad \nu = \lambda e, \\ ((Y_{-i}^{(\ell)})_S^T + (Z_{-i}^{(\ell)})_S^T)\nu = sgn(c_S). \end{array}\right\}$$

This construction of ν satisfies all conditions in Lemma 12 with respect to

$$\begin{cases} c_i = [0, ..., 0, c_i^{(\ell)}, 0, ..., 0] \text{ with } c_i^{(\ell)} = c, \\ e_i = e, \end{cases} \tag{5.8}$$

except

$$\left\|[X_1, ..., X_{\ell-1}, X_{\ell+1}, ..., X_L]^T\nu\right\|_\infty < 1,$$

i.e., we must check for all data point $x \in \mathcal{X} \setminus \mathcal{X}^\ell$,

$$|\langle x, \nu\rangle| < 1. \tag{5.9}$$

Thus, if we show that the solution of (5.7) ν also satisfies (5.9), we can conclude that ν is a dual certificate required in Lemma 12, which implies that the candidate solution (5.8) associated with optimal (c, e) of (5.6) is indeed the optimal solution of (5.1) and therefore SEP holds.

6. To be precise, it is the corresponding $c_i = [0, ..., 0, (c_i^{(\ell)})^T, 0, ..., 0]^T$ that satisfies SEP.

5.3 Dual separation condition

Our strategy to show (5.9) is to provide an upper bound of $|\langle x, \nu\rangle|$ then impose the inequality on the upper bound.

First, we find it appropriate to project ν to the subspace $\mathcal{S}_\ell$ and its orthogonal complement subspace then analyze separately. For convenience, denote $\nu_1 := \mathbb{P}_{S_\ell}(\nu)$, $\nu_2 := \mathbb{P}_{\mathcal{S}_\ell^\perp}(\nu)$. Then

$$\begin{aligned} |\langle x, \nu\rangle| =& |\langle y + z, \nu\rangle| \le |\langle y, \nu_1\rangle| + |\langle y, \nu_2\rangle| + |\langle z, \nu\rangle| \\ \le& \mu(\mathcal{X}_\ell)\|\nu_1\| + \|y\|\|\nu_2\||\cos(\angle(y, \nu_2))| + \|z\|\|\nu\||\cos(\angle(z, \nu))|. \end{aligned} \tag{5.10}$$

To see the last inequality, check that by Definition 3, $|\langle y, \frac{\nu_1}{\|\nu_1\|}\rangle| \le \mu(\mathcal{X}_\ell)$.

Since we are considering general (possibly adversarial) noise, we will use the relaxation $|\cos(\theta)| \le 1$ for all cosine terms (a better bound under random noise will be given later). Thus, what left is to bound $\|\nu_1\|$ and $\|\nu_2\|$ (note $\|\nu\| = \sqrt{\|\nu_1\|^2 + \|\nu_2\|^2} \le \|\nu_1\| + \|\nu_2\|$).

5.3.1 Bounding $\|\nu_1\|$

We first bound $\|\nu_1\|$ by exploiting the feasible region of ν_1 in (5.7):

$$\left\{\nu \middle| \|(X_{-i}^{(\ell)})^T\nu\|_\infty \le 1\right\},$$

which is equivalent to

$$\left\{\nu \middle| x_j^T\nu \le 1 \quad \text{for every column } x_j \text{ of } X_{-i}^{(\ell)}\right\}.$$

Decompose the condition into

$$y_j^T\nu_1 + (\mathbb{P}_{\mathcal{S}_\ell} z_j)^T\nu_1 + z_j^T\nu_2 \le 1.$$

and relax the expression into

$$y_j^T\nu_1 + (\mathbb{P}_{\mathcal{S}_\ell} z_j)^T\nu_1 \le 1 - z_j^T\nu_2 \le 1 + \delta\|\nu_2\|. \tag{5.11}$$

The relaxed condition contains the feasible region of ν_1 in (5.7). It turns out that the geometric properties of this relaxed feasible region provides an upper bound of $\|\nu_1\|$.

Definition 13 (polar set) *The polar set $\mathcal{K}^o$ of set $\mathcal{K} \in \mathbb{R}^d$ is defined as*

$$\mathcal{K}^o = \left\{y \in \mathbb{R}^d : \langle x, y\rangle \le 1 \text{ for all } x \in \mathcal{K}\right\}.$$

By the polytope geometry, we have

$$\|(Y_{-i}^{(\ell)} + \mathbb{P}_{\mathcal{S}_\ell}(Z_{-i}^{(\ell)}))^T\nu_1\|_\infty \le 1 + \delta\|\nu_2\| \;\Leftrightarrow\; \nu_1 \in \left[\mathcal{P}\left(\frac{Y_{-i}^{(\ell)} + \mathbb{P}_{\mathcal{S}_\ell}(Z_{-i}^{(\ell)})}{1 + \delta\|\nu_2\|}\right)\right]^o := \mathcal{T}^o. \tag{5.12}$$

Now we introduce the concept of circumradius.

Definition 14 (circumradius) *The circumradius of a convex body $\mathcal{P}$, denoted by $R(\mathcal{P})$, is defined as the radius of the smallest Euclidean ball containing $\mathcal{P}$.*

The magnitude $\|\nu_1\|$ is bounded by $R(\mathcal{T}^o)$. Moreover, by the the following lemma we may find the circumradius by analyzing the polar set of $\mathcal{T}^o$ instead. By the property of polar operator, polar of a polar set gives the tightest convex envelope of the original set, i.e., $(\mathcal{K}^o)^o = conv(\mathcal{K})$. Since $\mathcal{T} = \text{conv}\left(\pm\frac{Y_{-i}^{(\ell)}+\mathbb{P}_{\mathcal{S}_\ell}(Z_{-i}^{(\ell)})}{1+\delta\|\nu_2\|}\right)$ is convex in the first place, the polar set of $\mathcal{T}^o$ is $\mathcal{T}$.

Lemma 15 (Page 448 in Brandenberg et al. (2004)) *For a symmetric convex body $\mathcal{P}$, i.e. $\mathcal{P} = -\mathcal{P}$, inradius of $\mathcal{P}$ and circumradius of polar set of $\mathcal{P}$ satisfy:*

$$r(\mathcal{P})R(\mathcal{P}^o) = 1.$$

Lemma 16 *Given $X = Y + Z$, denote $\rho := \max_i \|\mathbb{P}_{\mathcal{S}} z_i\|$, furthermore $Y \in \mathcal{S}$ where $\mathcal{S}$ is a linear subspace, then we have:*

$$r(\text{Proj}_{\mathcal{S}}(\mathcal{P}(X))) \geq r(\mathcal{P}(Y)) - \rho$$

Proof First note that projection to a subspace is a linear operator. Hence $\text{Proj}_{\mathcal{S}}(\mathcal{P}(X)) = \mathcal{P}(\mathbb{P}_{\mathcal{S}}X)$. Then by definition, the boundary set of $\mathcal{P}(\mathbb{P}_{\mathcal{S}}X)$ is $\mathcal{B} := \{y \mid y = \mathbb{P}_{\mathcal{S}}Xc; \|c\|_1 = 1\}$. Inradius by definition is the largest ball containing in the convex body, hence $r(\mathcal{P}(\mathbb{P}_{\mathcal{S}}X)) = \min_{y\in\mathcal{B}}\|y\|$. Now we provide a lower bound of it:

$$\|y\| \geq \|Yc\| - \|\mathbb{P}_{\mathcal{S}}Zc\| \geq r(\mathcal{P}(Y)) - \sum\nolimits_j \|\mathbb{P}_{\mathcal{S}}z_j\||c_j| \geq r(\mathcal{P}(Y)) - \rho\|c\|_1.$$

This concludes the proof. ■

A bound of $\|\nu_1\|$ follows directly from Lemma 15 and Lemma 16:

$$\begin{aligned}\|\nu_1\| \leq& (1+\delta\|\nu_2\|)R(\mathcal{P}(Y_{-i}^{(\ell)} + \mathbb{P}_{\mathcal{S}_\ell}(Z_{-i}^{(\ell)})))\\ =& \frac{1+\delta\|\nu_2\|}{r(\mathcal{P}(Y_{-i}^{(\ell)} + \mathbb{P}_{\mathcal{S}_\ell}(Z_{-i}^{(\ell)}))} = \frac{1+\delta\|\nu_2\|}{r(\text{Proj}_{\mathcal{S}_\ell}(\mathcal{P}(X_{-i}^{(\ell)})))} \leq \frac{1+\delta\|\nu_2\|}{r\left(\mathcal{Q}_{-i}^{\ell}\right) - \delta_1}.\end{aligned} \tag{5.13}$$

This bound depends on $\|\nu_2\|$, which we analyze below.

5.3.2 Bounding $\|\nu_2\|$

Since ν is the optimal solution to $\mathbf{D}_1$, it obeys the second optimality condition in Lemma 12:

$$\nu = \lambda e_i = \lambda(x_i - X_{-i}^{(\ell)}c).$$

By projecting ν to $\mathcal{S}_\ell^\perp$, we get $\nu_2 = \lambda\mathbb{P}_{\mathcal{S}_\ell^\perp}(x_i - X_{-i}^{(\ell)}c) = \lambda\mathbb{P}_{\mathcal{S}_\ell^\perp}(z_i - Z_{-i}^{(\ell)}c)$. It follows that

$$\begin{aligned}\|\nu_2\| &\leq \lambda\left(\|\mathbb{P}_{\mathcal{S}_\ell^\perp}z_i\| + \|\mathbb{P}_{\mathcal{S}_\ell^\perp}Z_{-i}^{(\ell)}c\|\right)\\ &\leq \lambda\left(\|\mathbb{P}_{\mathcal{S}_\ell^\perp}z_i\| + \sum_j |c_j|\|\mathbb{P}_{\mathcal{S}_\ell^\perp}z_j\|\right)\\ &\leq \lambda(\|c\|_1+1)\delta_2 \leq \lambda(\|c\|_1+1)\delta.\end{aligned} \tag{5.14}$$

Now we bound $\|c\|_1$. Since (c, e) is the optimal solution, $\|c\|_1 + \frac{\lambda}{2}\|e\|^2 \leq \|\tilde{c}\|_1 + \frac{\lambda}{2}\|\tilde{e}\|^2$ for any feasible solution $(\tilde{c}, \tilde{e})$. Let $\tilde{c}$ be the solution of

$$\min_c \|c\|_1 \quad s.t. \quad y_i^{(\ell)} = Y_{-i}^{(\ell)} c, \tag{5.15}$$

then by strong duality,

$$\|\tilde{c}\|_1 = \max_\nu \left\{ \langle \nu, y_i^{(\ell)} \rangle \mid \|[Y_{-i}^{(\ell)}]^T \nu\|_\infty \leq 1 \right\}.$$

By Lemma 15, the optimal dual solution $\tilde{\nu}$ satisfies $\|\tilde{\nu}\| \leq \frac{1}{r(\mathcal{Q}_{-i}^\ell)}$. It follows that

$$\|\tilde{c}\|_1 = \langle \tilde{\nu}, y_i^{(\ell)} \rangle = \|\tilde{\nu}\| \|y_i^{(\ell)}\| \leq \frac{1}{r(\mathcal{Q}_{-i}^\ell)}.$$

On the other hand, $\tilde{e} = z_i - Z_{-i}^{(\ell)} \tilde{c}$, so $\|\tilde{e}\|^2 \leq (\|z_i\| + \sum_j \|z_j\| |\tilde{c}_j|)^2 \leq (\delta + \|\tilde{c}\|_1 \delta)^2$, thus

$$\|c\|_1 \leq \|\tilde{c}\|_1 + \frac{\lambda}{2}\|\tilde{e}\|^2 - \frac{\lambda}{2}\|e\|^2 \leq \frac{1}{r(\mathcal{Q}_{-i}^\ell)} + \frac{\lambda}{2}\delta^2 \left[1 + \frac{1}{r(\mathcal{Q}_{-i}^\ell)}\right]^2 - \frac{1}{2\lambda}\|\nu_2\|^2.$$

Note that we used the property $\frac{\lambda}{2}\|e\|^2 = \frac{1}{2\lambda}\|\nu\|^2 \geq \frac{1}{2\lambda}\|\nu_2\|^2$. Substitute the bound of $\|c_1\|_1$ into (5.14) we get

$$\begin{aligned} & \|\nu_2\| \leq \lambda \left(\frac{1}{r(\mathcal{Q}_{-i}^\ell)} + \frac{\lambda}{2}\delta^2 \left[1 + \frac{1}{r(\mathcal{Q}_{-i}^\ell)}\right]^2 + 1 \right) \delta - \frac{\delta}{2}\|\nu_2\|^2 \\ \Leftrightarrow \quad & \|\nu_2\| + \frac{\delta}{2}\|\nu_2\|^2 \leq \lambda\delta \left(\frac{1}{r(\mathcal{Q}_{-i}^\ell)} + 1 \right) + \frac{\delta}{2} \left[\lambda\delta \left(\frac{1}{r(\mathcal{Q}_{-i}^\ell)} + 1 \right) \right]^2. \end{aligned}$$

Since function $f(\alpha) = \alpha + \frac{\delta}{2}\alpha^2$ monotonically increases when $\alpha > 0$, the above inequality implies

$$\|\nu_2\| \leq \lambda\delta \left(\frac{1}{r(\mathcal{Q}_{-i}^\ell)} + 1 \right), \tag{5.16}$$

which gives the desired bound for $\|\nu_2\|$.

5.3.3 Conditions for $|\langle x, \nu \rangle| < 1$

Putting together (5.10), (5.13) and (5.16), we have the upper bound of $|\langle x, \nu \rangle|$:

$$\begin{aligned} & |\langle x, \nu \rangle| \leq (\mu(\mathcal{X}_\ell) + \|\mathbb{P}_{\mathcal{S}_\ell} z\|)\|\nu_1\| + (\|y\| + \|\mathbb{P}_{\mathcal{S}_\ell^\perp} z\|)\|\nu_2\| \\ \leq & \frac{\mu(\mathcal{X}_\ell) + \delta_1}{r(\mathcal{Q}_{-i}^\ell) - \delta_1} + \left(\frac{(\mu(\mathcal{X}_\ell) + \delta_1)\delta}{r(\mathcal{Q}_{-i}^\ell) - \delta_1} + 1 + \delta \right) \|\nu_2\| \\ \leq & \frac{\mu(\mathcal{X}_\ell) + \delta_1}{r(\mathcal{Q}_{-i}^\ell) - \delta_1} + \lambda\delta(1+\delta) \left(\frac{1}{r(\mathcal{Q}_{-i}^\ell)} + 1 \right) + \frac{\lambda\delta^2(\mu(\mathcal{X}_\ell) + \delta_1)}{r(\mathcal{Q}_{-i}^\ell) - \delta_1} \left(\frac{1}{r(\mathcal{Q}_{-i}^\ell)} + 1 \right). \end{aligned}$$

For convenience, we further relax the second $r(\mathcal{Q}^{\ell}_{-i})$ into $r(\mathcal{Q}^{\ell}_{-i}) - \delta_1$. The dual separation condition is thus guaranteed with

$$\frac{\mu(\mathcal{X}_\ell) + \delta_1 + \lambda\delta(1+\delta) + \lambda\delta^2(\mu(\mathcal{X}_\ell) + \delta_1)}{r\big(\mathcal{Q}^{\ell}_{-i}\big) - \delta_1} + \lambda\delta(1+\delta) + \frac{\lambda\delta^2(\mu(\mathcal{X}_\ell) + \delta_1)}{r\big(\mathcal{Q}^{\ell}_{-i}\big)(r\big(\mathcal{Q}^{\ell}_{-i}\big) - \delta_1)} < 1.$$

Denote $\rho := \lambda\delta(1+\delta)$, assume $\delta < r\big(\mathcal{Q}^{\ell}_{-i}\big)$, $(\mu(\mathcal{X}_\ell) + \delta_1) < 1$ and simplify the form with

$$\frac{\lambda\delta^2(\mu(\mathcal{X}_\ell) + \delta_1)}{r\big(\mathcal{Q}^{\ell}_{-i}\big) - \delta_1} + \frac{\lambda\delta^2(\mu(\mathcal{X}_\ell) + \delta_1)}{r\big(\mathcal{Q}^{\ell}_{-i}\big)(r\big(\mathcal{Q}^{\ell}_{-i}\big) - \delta_1)} < \frac{\rho}{r\big(\mathcal{Q}^{\ell}_{-i}\big) - \delta_1},$$

we get a sufficient condition

$$\mu(\mathcal{X}_\ell) + 2\rho + \delta_1 < (1-\rho)\,(r(\mathcal{Q}^{\ell}_{-i}) - \delta_1). \tag{5.17}$$

To generalize (5.17) to all data of all subspaces, the following must hold for each $\ell = 1, ..., k$:

$$\mu(\mathcal{X}_\ell) + 2\rho + \delta_1 < (1-\rho)\left(\min_{\{i:x_i \in X^{(\ell)}\}} r(\mathcal{Q}^{(\ell)}_{-i}) - \delta_1\right). \tag{5.18}$$

This gives a first condition on δ and λ (wihtin ρ), which we call it "**dual separation condition**" under noise. Note that this reduces to exactly the geometric condition in Soltanolkotabi and Candes (2012)'s Theorem 2.5 when $\delta = 0$.

5.4 Avoid trivial solutions

Besides SEP, we also need to show the solution is non-trivial. The idea is that when λ is large enough, the trivial solution $c^* = 0$, $e^* = x_i^{(\ell)}$ can never be optimal.

As we trace along the regularization path by increasing λ from 0, one column of the design matrix X_{-i} will enter the support set. This column will be the one that attains $\|X^T_{-i}x_i\|_\infty$, and $\lambda = \frac{1}{\|X^T_{-i}x_i\|_\infty}$ when it happens. Therefore, as long as $\lambda > \frac{1}{\|X^T_{-i}x_i\|_\infty}$, the solution will not be trivial.

Note that under the dual separation condition, we only need to consider points in the same subspace. So $\|X^T_{-i}x_i\|_\infty = \left\|[X^{(\ell)}_{-i}]^T x_i\right\|_\infty$. Let $x_j \in X^{(\ell)}_{-i}$ be the column that attains the maximum in $\left\|[X^{(\ell)}_{-i}]^T x_i\right\|_\infty$ and $y_k \in Y^{(\ell)}_{-i}$ be the column that attains the maximum in $\left\|[Y^{(\ell)}_{-i}]^T y_i\right\|_\infty$ (if there are more than one maximizers, pick any one), we can write

$$\begin{aligned}
\left\|[X^{(\ell)}_{-i}]^T x_i\right\|_\infty &= |\langle x_j, x_i\rangle| \geq |\langle x_k, x_i\rangle| \\
&= |\langle y_k, y_i\rangle + \langle y_k, z_i\rangle + \langle z_k, y_i\rangle + \langle z_k, z_i\rangle| \\
&\geq |\langle y_k, y_i\rangle| - |\langle y_k, z_i\rangle + \langle z_k, y_i\rangle + \langle z_k, z_i\rangle| \\
&= \left\|[Y^{(\ell)}_{-i}]^T y_i\right\|_\infty - |\langle y_k, z_i\rangle + \langle z_k, y_i\rangle + \langle z_k, z_i\rangle| \\
&\geq r(\mathcal{Q}^{(\ell)}_{-i}) - 2\delta - \delta^2.
\end{aligned} \tag{5.19}$$

The last inequality follows from the upper bound of noise magnitude and the observation that the inradius of $\mathcal{Q}_{-i}^{(\ell)}$ defines a uniform lower bound of $\left\|[Y_{-i}^{(\ell)}]^T w\right\|_\infty$ for any unit vector $w \in \mathcal{S}_\ell$. Therefore, as long as

$$\lambda \geq \frac{1}{r(\mathcal{Q}_{-i}^{(\ell)}) - 2\delta - \delta^2}, \tag{5.20}$$

the solution c_i for ith column is not trivial. This bound is strictly better than what we obtain in the conference version (Wang and Xu, 2013) and is the key for improving the rate for noise tolerance over the previous version. Also, check that

$$\delta < \frac{r(r_\ell - \mu_\ell)}{2 + 7r_\ell} \tag{5.21}$$

under bound of δ in the theorem statement, $r(\mathcal{Q}_{-i}^{(\ell)}) - 2\delta - \delta^2 > 0$ for any i, ℓ.

A side remark is that the Lasso regularization path is formally described in Tibshirani et al. (2013) and it is unique whenever the data points are in general position. As a result, we can potentially calculate the entry point of kth non-zero coefficient for any $0 < k < d$, any x_i and X_{-i}. This would however complicate the results unnecessarily, as Lasso path is not monotone (some coefficient may leave the support set as λ increases). We therefore stick to the simpler requirement of c_i being non-trivial.

5.5 Existence of a proper λ

Basically, (5.18) and (5.20) must be satisfied simultaneously for all $\ell = 1, ..., L$. Essentially (5.20) gives a condition of λ from below, and (5.18) gives a condition from above. Recall that the denotations $r_\ell := \min_{\{i:x_i \in X^{(\ell)}\}} r(\mathcal{Q}_{-i}^{(\ell)})$, $\mu_\ell := \mu(\mathcal{X}_\ell)$ and $r = \min_\ell r_\ell$, the condition on λ is:

$$\max_\ell \frac{1}{r_\ell - 2\delta - \delta^2} < \lambda < \min_\ell \frac{r_\ell - \mu_\ell - 2\delta_1}{\delta(1+\delta)(2 + r_\ell - \delta_1)}.$$

With the observation that

$$\max_\ell \frac{1}{r_\ell - 2\delta - \delta^2} = \frac{1}{\min r_\ell - 2\delta - \delta^2},$$

it suffices to require λ to obey for each ℓ:

$$\frac{1}{r - 2\delta - \delta^2} < \lambda < \frac{r_\ell - \mu_\ell - 2\delta_1}{\delta(1+\delta)(2 + r_\ell - \delta_1)}. \tag{5.22}$$

We will now show that under condition (5.21), the range (5.22) is not an empty set. Again, we relax δ_1 to δ in (5.22) and get

$$\frac{1}{r - 2\delta - \delta^2} < \frac{r_\ell - \mu_\ell - 2\delta}{\delta(1+\delta)(2 + r_\ell - \delta)}. \tag{5.23}$$

Since all denominators are positive, we obtain the standard form of the inequality

$$A\delta^3 + B\delta^2 + C\delta + D < 0$$

with

$$\begin{cases} A = -3 \leq 0 \\ B = -3 + 2(r_\ell - \mu_\ell) + r_\ell \leq 0 \\ C = 2 + 4(r_\ell - \mu_\ell) + r_\ell + 2r \leq 2 + 7r_\ell \\ D = -r(r_\ell - \mu_\ell) \end{cases}$$

Check that (5.21) is sufficient for the above 3rd order inequality to hold. Therefore,

$$(5.21) \Rightarrow A\delta^3 + B\delta^2 + C\delta + D < 0 \Leftrightarrow (5.23) \Rightarrow (5.22) \text{ is not an empty set.}$$

This completes the proof of Theorem 6.

6. Proof of Results for Randomized Cases

In this section, we provide proofs to the theorems of the three randomized models:

- **Deterministic data+random noise;**
- **Semi-random data+random noise;**
- **Fully random.**

To do this, we need to bound δ_1, $\cos(\angle(z, \nu))$ and $\cos(\angle(y, \nu_2))$ when Z follows the *Random Noise Model*, such that a better dual separation condition can be obtained. Moreover, for the *Semi-random* and the *Random data model*, we need to bound $r(\mathcal{Q}_{-i}^{(\ell)})$ when data samples from each subspace are drawn uniformly and bound $\mu(\mathcal{X}_\ell)$ when subspaces are randomly generated. These require the following lemmas.

Lemma 17 (Upper bound on the area of spherical cap) *Let $a \in \mathbb{R}^n$ be a random vector sampled from a unit sphere and z is a fixed vector. Then we have:*

$$Pr\left(|a^T z| > \epsilon \|z\|\right) \leq 2e^{\frac{-n\epsilon^2}{2}}$$

This Lemma is extracted from an equation in page 29 of Soltanolkotabi and Candes (2012), which is in turn adapted from the upper bound on the area of spherical cap in Ball (1997). By definition of the Random Noise Model, z_i is spherical symmetric, which implies that the direction of z_i is distributed uniformly on the n-dimensional unit sphere. Hence Lemma 17 applies whenever an inner product involves z. As an example, we write the following lemma.

Lemma 18 (Properties of Gaussian noise) *For Gaussian random matrix $Z \in \mathbb{R}^{n \times N}$, if each entry $Z_{i,j} \sim N(0, \frac{\sigma}{\sqrt{n}})$, then each column z_i satisfies:*

1. $Pr(\|z_i\|^2 > (1+t)\sigma^2) \leq e^{\frac{n}{2}(\log(t+1)-t)}$
2. $Pr(|\langle z_i, z\rangle| > \epsilon \|z_i\| \|z\|) \leq 2e^{\frac{-n\epsilon^2}{2}}$

where z is any fixed vector, or a random vector that is independent to z_i.

Proof The second property follows directly from Lemma 17 as Gaussian vector has a uniformly random direction.

To show the first property, we observe that the sum of n independent square Gaussian random variables follows χ^2 distribution with degree of freedom n. In other words, we have

$$\|z_i\|^2 = |Z_{1i}|^2 + ... + |Z_{ni}|^2 \sim \frac{\sigma^2}{n}\chi^2(n).$$

By Hoeffding's inequality, we have an approximation of its CDF (Dasgupta and Gupta, 2002), which gives us

$$Pr(\|z_i\|^2 > \alpha\sigma^2) = 1 - \mathrm{CDF}_{\chi^2_n}(\alpha) \leq (\alpha e^{1-\alpha})^{\frac{n}{2}}.$$

Substitute $\alpha = 1 + t$, we obtain the concentration statement in the lemma. ■

By Lemma 18, $\delta = \max_i \|z_i\|$ is bounded with high probability. δ_1 can be bounded even more tightly because each $\mathcal{S}_\ell$ is low-rank. Likewise, $\cos(\angle(z, \nu))$ is bounded by a small value with high probability. Moreover, since $\nu = \lambda e = \lambda(x_i - X_{-i}c)$, $\nu_2 = \lambda \mathbb{P}_{\mathcal{S}_\ell^\perp}(z_i - Z_{-i}c)$. Thus ν_2 is indeed a weighted sum of random noise in a $(n - d_\ell)$-dimensional subspace. Consider y a fixed vector, $\cos(\angle(y, \nu_2))$ is also bounded with high probability.

Replace these observations into (5.9) and the corresponding bound of $\|\nu_1\|$ and $\|\nu_2\|$, we obtain the equivalent *dual separation condition* under the random noise model (equivalent to (5.17) in the proof of the deterministic case). This is formalized in the following lemma.

Lemma 19 (Dual separation condition under random noise) *Let* $\rho := \lambda\delta(1+\delta)$ *and*

$$\epsilon := \sqrt{\frac{6\log N}{n - \max_\ell d_\ell}} \leq \sqrt{\frac{C\log(N)}{n}}$$

for some constant C*. Under random noise model, if for each* $\ell = 1, ..., L$

$$\mu(\mathcal{X}_\ell) + \delta\epsilon + 3\rho\epsilon \leq (1 - \rho\epsilon)(\max_i r(\mathcal{Q}_{-i}^{(\ell)}) - \delta\sqrt{d_\ell}\epsilon), \tag{6.1}$$

then dual separation condition (5.9) *holds for all data points with probability at least* $1 - 8/N$*.*

Proof Recall that we want to find an upper bound of $|\langle x, \nu\rangle|$.

$$|\langle x, \nu\rangle| \leq \mu\|\nu_1\| + \|y\|\|\nu_2\||\cos(\angle(y, \nu_2))| + \|z\|\|\nu\||\cos(\angle(z, \nu))| \tag{6.2}$$

Here we will bound the two cosine terms and δ_1 under the random noise model.

As discussed above, directions of z and ν_2 are independently and uniformly distributed on the n-dimension unit sphere. Then by Lemma 17,

$$\begin{cases} Pr\left(\cos(\angle(z, \nu)) > \sqrt{\frac{6\log N}{n}}\right) \leq \frac{2}{N^3}; \\ Pr\left(\cos(\angle(y, \nu_2)) > \sqrt{\frac{6\log N}{n - d_\ell}}\right) \leq \frac{2}{N^3}; \\ Pr\left(\cos(\angle(z, \nu_2)) > \sqrt{\frac{6\log N}{n}}\right) \leq \frac{2}{N^3}. \end{cases}$$

Using the same technique, we derive a bound for δ_1. Given an orthonormal basis U of S_ℓ, $\mathbb{P}_{S_\ell} z = UU^T z$, then

$$\|UU^T z\| = \|U^T z\| = \sqrt{\sum_{i=1,\ldots,d_\ell} |U_{:,i}^T z|^2}.$$

Apply Lemma 17 for each i, then by union bound, we get:

$$Pr\left(\|\mathbb{P}_{S_\ell} z\| > \sqrt{\frac{6d_\ell \log N}{n}}\delta\right) \leq \frac{2d_\ell}{N^3}.$$

Since δ_1 is the worse case bound for all L subspace and all N noise vector, then a union bound gives:

$$Pr\left(\delta_1 > \sqrt{\frac{6d_\ell \log N}{n}}\delta\right) \leq \frac{2\sum_\ell d_\ell}{N^2}$$

Moreover, we can find a probabilistic bound for $\|\nu_1\|$ too by a variation of (5.11) for the random case, which now becomes

$$y_i^T \nu_1 + (\mathbb{P}_{\mathcal{S}_\ell} z_i)^T \nu_1 \leq 1 - z_i^T \nu_2 \leq 1 + \delta_2 \|\nu_2\| |\cos \angle(z_i, \nu_2)|. \tag{6.3}$$

Substituting the upper bound of the cosines to (6.2) and (6.3), we get respectively

$$|\langle x, \nu\rangle| \leq \mu\|\nu_1\| + \|y\|\|\nu_2\|\sqrt{\frac{6\log N}{n-d_\ell}} + \|z\|\|\nu\|\sqrt{\frac{6\log N}{n}},$$

and

$$\|\nu_1\| \leq \frac{1 + \delta\|\nu_2\|\sqrt{\frac{6\log N}{n}}}{r(\mathcal{Q}_{-i}^\ell) - \delta_1}.$$

This new bound of $\|\nu_1\|$ follows from (6.3), Lemma 15 and 16. For the bound of $\|\nu_2\|$ we simply use (5.16):

$$\|\nu_2\| \leq \lambda\delta\left(\frac{1}{r(\mathcal{Q}_{-i}^\ell)} + 1\right).$$

To lighten notations in this proof, denote

$$r := r(\mathcal{Q}_{-i}^\ell), \quad \epsilon := \sqrt{\frac{6\log N}{n - \max_\ell d_\ell}}, \quad \mu := \mu(\mathcal{X}_\ell).$$

Substitute them in the bound, we get

$$\begin{aligned}
|\langle x, \nu\rangle| \leq & \frac{\mu + \delta\epsilon}{r - \epsilon\sqrt{d_\ell}\delta} + \frac{\lambda\delta^2(\mu + \delta\epsilon)\epsilon}{r - \epsilon\sqrt{d_\ell}\delta}\left(\frac{1}{r} + 1\right) + \lambda\delta\epsilon\left(\frac{1}{r} + 1\right) + \lambda\delta^2\epsilon\left(\frac{1}{r} + 1\right) \\
= & \frac{\mu + \delta\epsilon}{r - \epsilon\sqrt{d_\ell}\delta} + \frac{\lambda\epsilon\delta^2(\frac{\mu+\delta\epsilon}{r}) + \lambda\epsilon\delta^2(\mu + \delta\epsilon)}{r - \epsilon\sqrt{d_\ell}\delta} + \frac{\lambda\delta(\delta+1)\epsilon}{r} + \lambda\delta(\delta+1)\epsilon \\
\underset{*}{\leq} & \frac{\mu + \delta\epsilon}{r - \epsilon\sqrt{d_\ell}\delta} + \frac{\lambda\epsilon\delta^2}{r - \epsilon\sqrt{d_\ell}\delta} + \frac{\lambda\epsilon\delta^2}{r - \epsilon\sqrt{d_\ell}\delta} + \frac{\lambda\epsilon(\delta + \delta^2)}{r - \epsilon\sqrt{d_\ell}\delta} + \lambda\epsilon(\delta + \delta^2) \\
= & \frac{\mu + \delta\epsilon + \lambda\epsilon(\delta + 3\delta^2)}{r - \epsilon\sqrt{d_\ell}\delta} + \lambda\epsilon(\delta + \delta^2) \underset{**}{\leq} \frac{\mu + \delta\epsilon + 3\rho\epsilon}{r - \epsilon\sqrt{d_\ell}\delta} + \rho\epsilon.
\end{aligned} \tag{6.4}$$

In the inequality "$*$", we used $(\mu + \epsilon\delta)/r < 1$ and $\mu + \epsilon\delta < 1$; and in the inequality "$**$", we used $\lambda\delta^2 \leq \lambda\epsilon(\delta + \delta^2)$ and replaced all such expression with $\rho\epsilon$ that we defined earlier.

Now impose the dual detection constraint on the upper bound, we get:

$$\rho\epsilon + \frac{\mu + \delta\epsilon + 3\rho\epsilon}{r - \delta\sqrt{d_\ell}\epsilon} < 1.$$

Reorganized the inequality, we reach the desired condition:

$$\mu + \delta\epsilon < (1 - \rho\epsilon)(r - \delta\sqrt{d_\ell}\epsilon) - 3\rho\epsilon.$$

There are N^2 instances for each of the three events related to the consine value, apply union bound we get the failure probability $\frac{6}{N} + \frac{2\sum_\ell d_\ell}{N^2} \leq \frac{8}{N}$. Note $\sum_\ell d_\ell \leq N$ because one needs at least d_ℓ data points to span an d_ℓ dimensional subspace. This concludes the proof. ■

Lemma 20 (Avoid trivial solution under random noises) *Let* $\epsilon = \sqrt{\frac{6\log N}{n}}$ *and assume* $\min_\ell r_\ell - 2\epsilon\delta - \epsilon\delta^2 > 0$. *If we take*

$$\lambda > (r_\ell - 2\epsilon\delta - \epsilon\delta^2)^{-1} \tag{6.5}$$

for every $\ell = 1, ..., n$, *then the solution* $c_i \neq 0$ *for all* i *with probability at least* $1 - 6/N^2$.

Proof We use the same argument as in Section 5.4, except that we now have a tighter probabilistic bound for (5.19). For any i, k in the equation, z_i and z_k are independent to each other and to y_k, y_i respectively. Therefore, we can invoke Lemma 17 and obtain

$$|\langle y_k, z_i\rangle + \langle z_k, y_i\rangle + \langle z_k, z_i\rangle| \leq 2\epsilon\delta + \epsilon\delta^2,$$

with probability greater than $2/N^3$. The proof is complete by taking the union bound over all $3\sum_i N_\ell = 3N$ instances. ■

6.1 Proof of Theorem 8 for Deterministic Data and Random Noise

We now prove Theorem 8. Lemma 19 has already provided the separation condition. The things left are to find the range of λ and update the condition of δ.

The range of λ: The range of valid λ for the random noise case can be obtained by substituting $\delta_1 < \delta\sqrt{d_\ell}\epsilon$ in (6.5) and rewriting (6.1) with respect to λ. This gives us

$$\frac{1}{r - 2\epsilon\delta - \epsilon\delta^2} < \lambda < \frac{r_\ell - \mu_\ell - \delta\epsilon - \delta\sqrt{d_\ell}\epsilon}{\epsilon\delta(1 + \delta)(3 + r_\ell - \delta\sqrt{d_\ell}\epsilon)}. \tag{6.6}$$

We remark that acritical difference from the deterministic noise model is that now there is a small ϵ in the denominator of the upper endpoint of the interval. Assume small μ, the valid range of λ expands to an order of $\Theta(1/r) \leq \lambda \leq \Theta(r/(\epsilon\max\{\delta^2, \delta\}))$.

The condition of δ: Now we will show that the two conditions

$$\epsilon\delta < \min_\ell \frac{r_\ell - \mu_\ell}{2\sqrt{d_\ell}+2}, \qquad \text{and} \qquad \epsilon\delta(1+\delta) < \min_\ell \frac{r(r_\ell-\mu_\ell)}{4r_\ell+6},$$

stated in the Theorem 8 are sufficient for the three inequalities

$$\begin{cases} r_\ell - \mu_\ell - \delta\epsilon > 0; & (6.7)\\ r - 2\delta\epsilon - \epsilon\delta^2 > 0; & (6.8)\\ \dfrac{1}{r-2\epsilon\delta-\epsilon\delta^2} < \dfrac{r_\ell-\mu_\ell-\delta\epsilon-\delta\sqrt{d_\ell}\epsilon}{\epsilon\delta(1+\delta)(3+r_\ell-\delta\sqrt{d_\ell}\epsilon)}; & (6.9)\end{cases}$$

to hold for $\ell = 1, ..., L$. Note that we used (6.7) in (6.4) when we derive the dual separation condition and (6.8) is assumed in Lemma 20, lastly (6.9) ensures a valid λ to exist in (6.6). Inequality (6.7) and (6.8) hold trivially given the two conditions, it remains to show (6.9):

$$\begin{aligned}
\delta(1+\delta) < \frac{r(r_\ell-\mu_\ell)}{\epsilon(4r_\ell+6)} &\Leftrightarrow \epsilon\delta(1+\delta)(2r_\ell+3) < \frac{r(r_\ell-\mu_\ell)}{2}\\
&\Rightarrow \epsilon\delta(1+\delta)(r_\ell-\mu_\ell+r_\ell+3) < \frac{r(r_\ell-\mu_\ell)}{2}\\
&\Leftrightarrow \epsilon\delta(1+\delta)(r_\ell+3) + 2\epsilon\delta(1+\delta)\frac{r_\ell-\mu_\ell}{2} < \frac{r(r_\ell-\mu_\ell)}{2}\\
&\Leftrightarrow \frac{1}{r-2\epsilon\delta-2\epsilon\delta^2} < \frac{r_\ell-\mu_\ell}{2\epsilon\delta(1+\delta)(r_\ell+3)}\\
&\Rightarrow \frac{1}{r-2\epsilon\delta-\epsilon\delta^2} < \frac{r_\ell-\mu_\ell}{2\epsilon\delta(1+\delta)(r_\ell+3-\delta\sqrt{d_\ell}\epsilon)} \overset{(a)}{\Rightarrow} (6.9),
\end{aligned}$$

where (a) holds by applying the first condition. This concludes the proof for Theorem 8.

6.2 Proof of Theorem 10 for the Semirandom Model with Random Noise

To prove Theorem 10, we only need to bound the inradii r and the incoherence parameter μ under the new assumptions, then plug them into Theorem 8.

Lemma 21 (Inradius bound of random samples) *In random sampling setting, when each subspace is sampled $N_\ell = \kappa_\ell d_\ell$ data points randomly, we have:*

$$Pr\left\{ c(\kappa_\ell)\sqrt{\frac{\beta\log(\kappa_\ell)}{d_\ell}} \le r(\mathcal{Q}_{-i}^{(\ell)}) \text{ for all pairs } (\ell,i) \right\} \ge 1 - \sum_{\ell=1}^{L} N_\ell e^{-d_\ell^\beta N_\ell^{1-\beta}}$$

This is extracted from Section-7.2.1 of Soltanolkotabi and Candes (2012). $\kappa_\ell = (N_\ell - 1)/d_\ell$ is the relative number of iid samples. $c(\kappa)$ is some positive value for all $\kappa > 1$ and for a numerical value κ_0, if $\kappa > \kappa_0$, we can take $c(\kappa) = \frac{1}{\sqrt{8}}$. Take $\beta = 0.5$, we get the required bound of r in Theorem 10.

Now, we provide a probabilistic upper bound of the projected subspace incoherence condition under the semi-random model by adapting Lemma 7.5 of Soltanolkotabi and Candes (2012) into our new setup.

Lemma 22 (Incoherence bound) *In deterministic subspaces/random sampling setting, the subspace incoherence is bounded from above:*

$$Pr\Big\{\mu(\mathcal{X}_\ell) \le t\left(\log[(N_\ell+1)N_{\ell'}] + \log L\right)\frac{\mathrm{aff}(S_\ell, S_{\ell'})}{\sqrt{d_\ell}\sqrt{d_{\ell'}}} \\ \textit{for all pairs}(\ell,\ell') \textit{ with } \ell \neq \ell'\Big\} \ge 1 - \frac{1}{L^2}\sum_{\ell\neq\ell'}\frac{1}{(N_\ell+1)N_{\ell'}}e^{-\frac{t}{4}}.$$

Proof The proof is an extension of a similar proof in Soltanolkotabi and Candes (2012). First we will show that when noise $z_i^{(\ell)}$ is spherical symmetric, and clean data points $y_i^{(\ell)}$ has iid uniform random direction, projected dual directions $v_i^{(\ell)}$ also follows a uniform random distribution.

Now we prove the claim. First by definition,

$$v_i^{(\ell)} = v(x_i^{(\ell)}, X_{-i}^{(\ell)}, \mathcal{S}_\ell, \lambda) = \frac{\mathbb{P}_{S_\ell}\nu}{\|\mathbb{P}_{S_\ell}\nu\|} = \frac{\nu_1}{\|\nu_1\|}.$$

Recall that ν is the unique optimal solution of $\mathbf{D}_1$ (5.7). Fix λ, $\mathbf{D}_1$ depends on two inputs, so we denote $\nu(x, X)$ and consider ν a function. Moreover, $\nu_1 = \mathbb{P}_{\mathcal{S}}\nu$ and $\nu_2 = \mathbb{P}_{\mathcal{S}^\perp}\nu$. Let $U \in n \times d$ be a set of orthonormal basis of d-dimensional subspace $\mathcal{S}$ and a rotation matrix $R \in \mathbb{R}^{d\times d}$. Then rotation matrix within subspace is hence URU^T. Let

$$\begin{aligned} x_1 :=& \mathbb{P}_{\mathcal{S}}x = y + z_1 \sim URU^T y + URU^T z_1, \\ x_2 :=& \mathbb{P}_{\mathcal{S}^\perp}x = z_2. \end{aligned}$$

As y is distributed uniformly on the unit sphere of $\mathcal{S}$, and z is a spherical symmetric noise (hence z_1 and z_2 are also spherical symmetric in subspace), for any fixed $\|x_1\|$, the distribution is uniform on the sphere, namely the conditional distribution $Pr\left(x_1 | \|x_1\| = \alpha\right)$ is uniform on the sphere with radius α. It suffices to show that with fixed $\|x_1\|$, v (the unit direction of projected dual variable ν_1) also follows a uniform distribution on a unit sphere of the subspace.

Since inner product $\langle x, \nu\rangle = \langle x_1, \nu_1\rangle + \langle x_2, \nu_2\rangle$, we argue that if ν is the optimal solution of

$$\max_\nu \langle x, \nu\rangle - \frac{1}{2\lambda}\nu^T\nu, \quad \text{subject to:} \quad \|X^T\nu\|_\infty \le 1,$$

then the optimal solution of the following optimization

$$\begin{aligned} &\max_\nu \langle URU^T x_1 + x_2, \nu\rangle - \frac{1}{2\lambda}\nu^T\nu, \\ &\text{subject to:} \quad \|(URU^T X_1 + X_2)^T\nu\|_\infty \le 1, \end{aligned}$$

is indeed the transformed ν under the same R, i.e.,

$$\begin{aligned} \nu(R) &= \nu(URU^T x_1 + x_2, URU^T X_1 + X_2) \\ &= URU^T\nu_1(x, X) + \nu_2(x, X) = URU^T\nu_1 + \nu_2. \end{aligned} \tag{6.10}$$

To verify the argument, check that $\nu^T\nu = \nu(R)^T\nu(R)$ and

$$\langle URU^T x_1 + x_2, \nu(R)\rangle = \langle URU^T x_1, URU^T\nu_1\rangle + \langle x_1, \nu_2\rangle = \langle x, \nu\rangle$$

for all inner products in both objective function and constraints, preserving the optimality.

By projecting (6.10) to subspace, we show that operator $v(x, X, S)$ is linear *vis a vis* subspace rotation URU^T, i.e.,

$$v(R) = \frac{\mathbb{P}_{S_\ell}\nu(R)}{\|\mathbb{P}_{S_\ell}\nu(R)\|} = \frac{URU^T\nu_1}{\|URU^T\nu_1\|} = URU^T v. \tag{6.11}$$

On the other hand, we know that

$$URU^T x_1 + x_2 \sim x_1 + x_2, \qquad\qquad URU^T X_1 + X_2 \sim X_1 + X_2,$$

where $A \sim B$ means that the random variables A and B follows the same distribution. This is because when $\|x_1\|$ is fixed and each columns in X_1 has fixed magnitudes, $URU^T x_1 \sim x_1$ and $URU^T X_1 \sim X_1$. Also, adding additional random variables x_2 and X_2 changes the distribution the same way on both sides. Therefore,

$$v(R) = v(URU^T x_1 + x_2, URU^T X_1 + X_2, \mathcal{S}) \sim v(x, X, \mathcal{S}). \tag{6.12}$$

Combining (6.11) and (6.12), we conclude that for any rotation R

$$v_i^{(\ell)}(R) \sim URU^T v_i^{(\ell)}.$$

In other words, the distribution of $v_i^{(\ell)}$ is uniform on the unit sphere of $\mathcal{S}_\ell$.

After this key step, the rest is identical to the proof of Lemma 7.5 of Soltanolkotabi and Candes (2012). The idea is to use Lemma 17 (upper bound of area of spherical caps) to provide a probabilistic bound of the pairwise inner product and Borell's inequality to show the concentration around the expected cosine canonical angles, namely, $\|U^{(k)^T}U^{(\ell)}\|_F/\sqrt{d_\ell}$. The proof is standard so we omit it in this paper. ■

6.3 Proof of Theorem 11 for the Fully Random Model with Gaussian Noises

The proof of Theorem 11 essentially applies Theorem 8 with specific inradii bound and incoherence bound. The bound for inradius is given in Lemma 21 and we use the following Lemma extracted from Step 2 of Section 7.3 in Soltanolkotabi and Candes (2012) to bound the subspace incoherence.

Lemma 23 (Incoherence bound of random subspaces) *In random subspaces setting, the projected subspace incoherence is bounded from above:*

$$Pr\left\{\mu(\mathcal{X}_\ell) \le \sqrt{\frac{6\log N}{n}} \text{ for all } \ell\right\} \ge 1 - \frac{2}{N}.$$

Now that we have shown that the projected dual directions are randomly distributed in their respective subspace, together with the fact that the subspaces themselves are randomly generated, we conclude that all clean data points y and projected dual direction v from different subspaces can be considered iid generated from the ambient space. The proof of Lemma 23 follows by simply applying Lemma 17 and a union bound across all N^2 events.

7. Experiments

To demonstrate the practical implications of our robustness guarantees for LASSO-SSC, we conduct four numerical experiments including three with synthetic generated data and one with real data. For fast computation, we use ADMM implementation of LASSO solver[7] with default numerical parameters. Its complexity is proportional to the problem size and the convergence guarantee (Boyd et al., 2011). We also implement a simple ADMM solver for the matrix version SSC

$$\min_{C} \ \|C\|_1 + \frac{\lambda}{2}\|X - XC\|_F^2 \quad \text{s.t.} \quad \operatorname{diag}(C) = 0, \tag{7.1}$$

which is consistently faster than the column-by-column LASSO version. This algorithm is first described in Elhamifar and Vidal (2013). To be self-contained, we provide the pseudocode and some numerical simulation in the appendix.

7.1 Numerical simulation

Our three numerical simulations test the effects of noise magnitude δ, subspace rank d and number of subspace L respectively.

Methods: To test our methods for all parameters, we scan through an exponential grid of λ ranging from $\sqrt{n} \times 10^{-2}$ to $\sqrt{n} \times 10^{3}$. In all experiments, ambient dimension $n = 100$, relative sampling $\kappa = 5$, subspace and data are drawn uniformly at random from unit sphere and then corrupted by Gaussian noise $Z_{ij} \sim N(0, \sigma/\sqrt{n})$. We measure the success of the algorithm by the relative violation of Self-Expressiveness Property defined below.

$$\text{RelViolation}\,(C, \mathcal{M}) = \frac{\sum_{(i,j)\notin\mathcal{M}} |C|_{i,j}}{\sum_{(i,j)\in\mathcal{M}} |C|_{i,j}}$$

where $\mathcal{M}$ is the ground truth mask containing all (i, j) such that $x_i, x_j \in \mathcal{X}^{(\ell)}$ for some ℓ. Note that $\text{RelViolation}\,(C, \mathcal{M}) = 0$ implies that SEP is satisfied. We also check that there is no all-zero columns in C, and the solution is considered trivial otherwise.

Results: The simulation results confirm our theoretical findings. In particular, Figure 8 shows that LASSO subspace detection property is possible for a very large range of λ and the dependence on noise magnitude is roughly $1/\sigma$ as predicted in (4.1). In addition, the sharp contrast of Figure 10 and 11 demonstrates our observations on the sensitivity of d and L.

7.2 Face Clustering Experiments

In this section, we evaluate the noise robustness of with LASSO-SSC on Extended YaleB (Lee et al., 2005), a real life face dataset of 38 subjects. For each subject, 64 face images are taken under different illuminations.

Subspace Modeling of Face Images: According to Basri and Jacobs (2003), face images under different illuminations can be well-approximated by a 9-dimensional linear subspace. In addition, Zhou et al. (2007) reveals the underlying 3-dimensional subspace

7. Freely available at:
http://www.stanford.edu/~boyd/papers/admm/

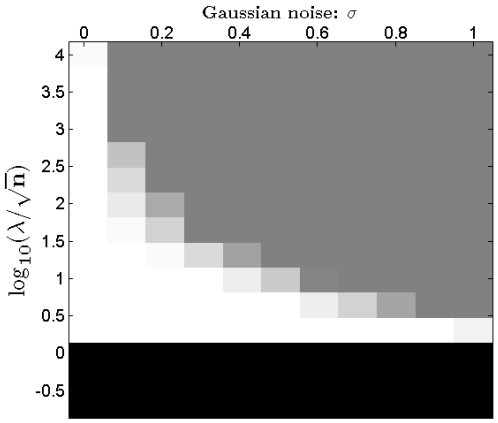


Figure 8: Exact recovery under noise. Simulated with $n = 100, d = 4, L = 3, \kappa = 5$ with increasing Gaussian noise $N(0, \sigma/\sqrt{n})$. **Black:** trivial solution ($C = 0$); **Gray:** RelViolation > 0.1; **White:** RelViolation $= 0$.

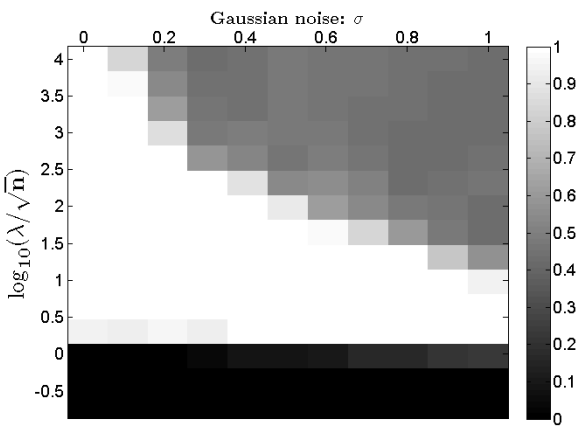


Figure 9: Spectral clustering accuracy for the experiment in Figure 8. The rate of accurate classification is represented in grayscale. White region means perfect classification. It is clear that exact subspace detection property (Definition 1) is not necessary for perfect classification.

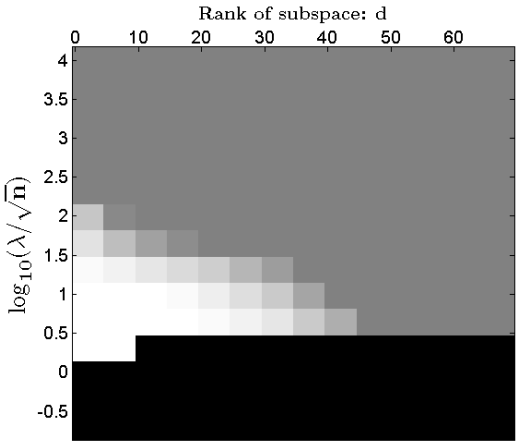


Figure 10: Effects of cluster rank d. Simulated with $n = 100, L = 3, \kappa = 5, \sigma = 0.2$ with increasing d. **Black:** trivial solution ($C = 0$); **Gray:** RelViolation > 0.1; **White:** RelViolation $= 0$. Observe that beyond a point, subspace detection property is not possible for any λ.

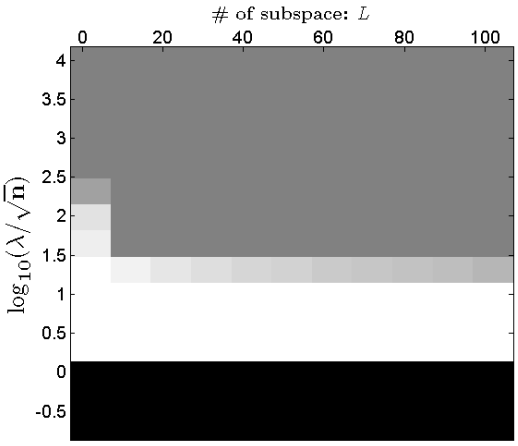


Figure 11: Effects of number of subspace L. Simulated with $n = 100, d = 2, \kappa = 5, \sigma = 0.2$ with increasing L. **Black:** trivial solution ($C = 0$); **Gray:** RelViolation > 0.1; **White:** RelViolation $= 0$. Note that even at the point when $dL = 200$(subspaces are highly dependent), subspace detection property holds for a large range of λ.

structure by assuming Lambert's reflectance and ignoring the shadow pixels. For the physics

of this subspace model, we refer the readers to Basri and Jacobs (2003) and Zhou et al. (2007) for detailed explanations.

Method: We conduct face clustering experiments on Extended YaleB Dataset with both 9D and 3D representations of face images and compare them under varying number of subspaces L and different level of injected noise. Specifically, the 9D subspaces are generated by projecting the data matrix corresponding to each subject to a 9D subspace via PCA and the 3D subspaces are generated by a factorization-based robust matrix completion method. Then we scan through a random selection of $[2, 4, 6, 10, 12, 18, 38]$ subjects and for each experiment we inject additive Gaussian noise $N(0, \sigma/\sqrt{n})$ with $\sigma = [0, 0.01, ..., 0.99, 1]$[8]. Each photo is resized to 48×42 so we have ambient dimension $n = 2016$ and there are 64 sample points for each subspace, hence $N = 64L$. The parameter λ is not carefully tuned, but simply chosen to be $\sqrt{n}$, which is order-wise correct for small σ according to (4.1).

Results: As we can see in Figure 12 and 13, the range where LASSO-Subspace Detection Property holds is much larger for the rank-3 experiments than the rank-9 experiments. Also, the recovery is not sensitive to the number of faces we want to cluster. Indeed, LASSO-SSC is able to succeed for both rank-9 and rank-3 data with a considerable range of noise even for the full 38 subjects clustering task.

These observations confirm our theoretical and simulation results—on deterministic subspace data from a real-life problem—that noise robustness of LASSO-SSC is sensitive to the subspace dimension d but not the number of subspaces L.

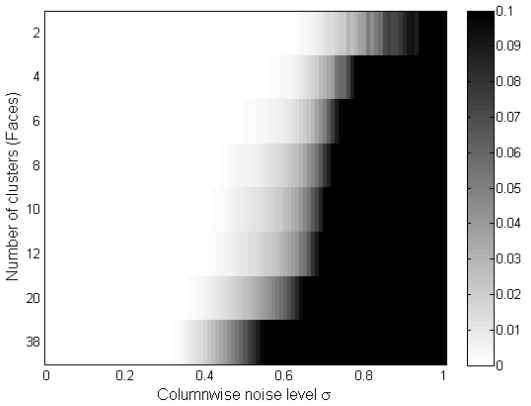


Figure 12: RelViolation of the Face clustering experiments with Rank 3 photometric face.

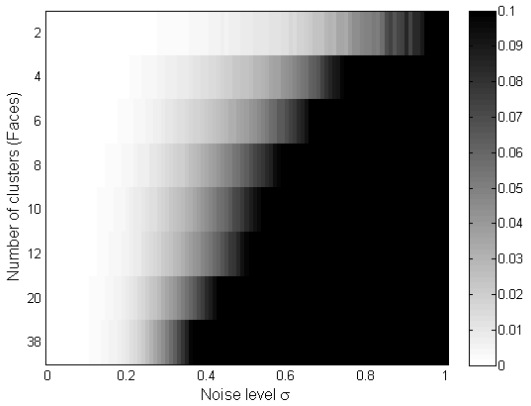


Figure 13: RelViolation of the Face clustering experiments with Rank 9 faces (after projection).

8. Conclusion and Future Directions

We presented a theoretical analysis for noisy subspace segmentation problem that is of great practical interests. We showed that the popular SSC algorithm *exactly* (not approximately)

8. In order to compare the effect of varying level of noise, we choose to inject artificial noise in this experiment. The performance of LASSO-SSC on real noise/data corruptions is well-documented in the motion segmentation experiments of Elhamifar and Vidal (2009, 2013)

detects the subspaces in the noisy case, which justified its empirical success on real problems. Our results are the first in showing LASSO-SSC to work under deterministic data and noise. For stochastic noise, we show that LASSO-SSC works even when noise is much larger than the signal. In addition, we discovered the orderwise relationship between LASSO-SSC's robustness to the level of noise and the subspace dimension, and we found that robustness is insensitive to the number of subspaces. These results lead to new theoretical understanding of SSC, and provide guidelines for practitioners and application-level researchers to judge whether SSC could possibly work well for their respective applications.

Open problems for subspace clustering include the graph connectivity problem raised by Nasihatkon and Hartley (2011), missing data problem (a first attempt is performed in Eriksson et al. (2012), which requires an unrealistic number of data), sparse corruptions on data and others. One direction closely related to this paper is to introduce a more practical metric of success. As we illustrated in the paper, subspace detection property is not necessary for perfect clustering. In fact from a pragmatic point of view, even perfect clustering is not necessary. Typical applications allow for a small number of misclassifications. It would be interesting to see whether stronger robustness results can be obtained for a more practical metric of success.

Acknowledgments

H. Xu was supported by the Ministry of Education of Singapore through AcRF Tier Two grant R-265-000-443-112.

Appendix A. Differences to Soltanolkotabi et al. (2014)

As we reviewed above, Soltanolkotabi et al. (2014) analyzed almost the same algorithm under the semi-random model. Besides the comparisons we made in Section 2 regarding the model of analysis and allowable noise level, there are a few other minor differences which we list here.

"Non-trivial" v.s. "Many true discoveries". In LASSO-Subspace Detection Property, we only require the resulting regression coefficient to be non-zero, while Soltanolkotabi et al. (2014, Theorem 3.2) has a result showing the conditions under which the number of non-zero coefficient is a constant fraction of subspace dimension d. Our results are weaker but more general (works without the semi-random assumption).

In fact, the conditions are more similar than different. We both pick regression coefficient of the same order in the semi-random model. In addition, when d becomes smaller than $\log \rho(i)/c_0$, these two conditions are essentially the same.

Choosing parameter λ. Soltanolkotabi et al. (2014) provides a two-pass mechanism to adaptively tune the parameter λ for each Lasso-SSC and the results are proven for this particular λ. On the other hand, our results are stated for any λ in a specified range. The choice of λ can also be independently tuned for each Lasso-SSC.

In practice, it is advisable to choose λ slightly larger than what is required for it to be non-trivial. We described two strategies in the discussion underneath Theorem 6.

Proof techniques. The proofs are admittedly similar in many ways (since we solve the same problem!), but the key technical components in controlling the magnitude of dual variables ν_1 and ν_2 are different. Soltanolkotabi et al. (2014, Lemma 8.5) relies on the semi-random model, and the resulting restricted isometry property (of the block of data points corresponding to each subspace). In contrast, our bound for $\|\nu_2\|$ does not require any probabilistic assumptions, therefore more general. It is however looser than Soltanolkotabi et al. (2014, Lemma 8.5) by a factor of $\sqrt{d}$ when we specialized in the semi-random model. This is probably what led to our worse dependence on the subspace dimension d in the bound we described in the discussion of Theorem 10.

In conclusion, we find that our results are complementary to that in Soltanolkotabi et al. (2014) and provide a novel point of view to the theoretical analysis for subspace clustering problems.

Appendix B. Numerical algorithm to solve Matrix-LASSO-SSC

In this section we outline the steps of solving the matrix version of LASSO-SSC below. Note that Elhamifar and Vidal (2012) derived a more general version of Matrix-LASSO-SSC to account for not only noisy but also sparse corruptions. We include this appendix merely for the convenience of the readers. Consider

$$\min_C \|C\|_1 + \frac{\lambda}{2}\|X - XC\|_F^2 \quad \text{s.t.} \quad \operatorname{diag}(C) = 0. \tag{B.1}$$

While this convex optimization problem can be solved by some off-the-shelf general purpose solvers such as SeDuMi (Sturm, 1999) or SDPT3 (Toh et al., 1999), such approaches are usually slow and non-scalable. An ADMM (Boyd et al., 2011) version of the problem is described here for fast computation. It solves an equivalent optimization program

$$\begin{aligned} &\min_C \|C\|_1 + \frac{\lambda}{2}\|X - XJ\|_F^2 \\ &\text{s.t.} \quad J = C - \operatorname{diag}(C). \end{aligned} \tag{B.2}$$

We add to the Lagrangian with an additional quadratic penalty term for the equality constraint and get the augmented Lagrangian

$$\mathcal{L} = \|C\|_1 + \frac{\lambda}{2}\|X - XJ\|_F^2 + \frac{\mu}{2}\|J - C + \operatorname{diag}(C)\|_F^2 + tr(\Lambda^T(J - C + \operatorname{diag}(C))),$$

where Λ is the dual variable and μ is a parameter. Optimization is done by alternatingly optimizing over J, C and Λ until convergence. The update steps are derived by solving $\partial\mathcal{L}/\partial J = 0$ and $\partial\mathcal{L}/\partial C = 0$. Notice that the objective function is non-differentiable for C at origin so we use the now standard soft-thresholding operator (Donoho, 1995). For both variables, the solution is given in closed-forms. For the update of Λ, we simply use the gradient descent method. For details of the ADMM algorithm and its guarantee, please refer to Boyd et al. (2011). To accelerate the convergence, it is possible to introduce a parameter ρ and increase μ by $\mu = \rho\mu$ at every iteration. The full algorithm is summarized in Algorithm 1.

Algorithm 1 Matrix-LASSO-SSC

Input: Data points as columns in $X \in \mathbb{R}^{n\times N}$, tradeoff parameter λ, numerical parameters μ_0 and ρ.
Initialize $C = 0$, $J = 0$, $\Lambda = 0$, $k = 0$.
while not converged **do**

1. Update J by
$$J = (\lambda X^T X + \mu_k I)^{-1}(\lambda X^T X + \mu_k C - \Lambda).$$
2. Update C by
$$C^{'} = \text{SoftThresh}_{\frac{1}{\mu_k}}\left(J + \Lambda/\mu_k\right),$$
$$C = C^{'} - \text{diag}(C^{'}).$$
3. Update Λ by
$$\Lambda = \Lambda + \mu_k(J - C)$$
4. Update parameter $\mu_{k+1} = \rho\mu_k$.
5. Iterate $k = k + 1$;

end while
Output: Affinity matrix $W = |C| + |C|^T$

Note that for the special case when $\rho = 1$, the inverse of $(\lambda Y^T Y + \mu I)$ can be precomputed, and hence each iteration can be computed in linear time. Empirically, we found it good to set $\mu = \lambda$ and it takes roughly 50-100 iterations to converge to a sufficiently good points. We remark that the matrix version of the algorithm is much faster than the column-by-column ADMM-Lasso and achieves almost the same numerical accuracy; see our experiments in Figure 14,15,16 and 17.

References

K. Ball. An elementary introduction to modern convex geometry. *Flavors of geometry*, 31: 1–58, 1997.

R. Basri and D.W. Jacobs. Lambertian reflectance and linear subspaces. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 25(2):218–233, 2003.

A. Ben-Tal and A. Nemirovski. Robust convex optimization. *Mathematics of Operations Research*, 23(4):769–805, 1998.

D. Bertsimas and M. Sim. The price of robustness. *Operations research*, 52(1):35–53, 2004.

S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

P.S. Bradley and O.L. Mangasarian. k-plane clustering. *Journal of Global Optimization*, 16 (1):23–32, 2000.

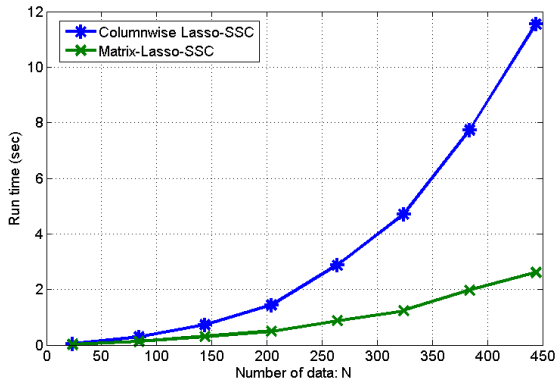


Figure 14: Run time comparison with increasing number of data. Simulated with $n = 100, d = 4, L = 3, \sigma = 0.2$, κ increases from 2 to 40 such that the number of data goes from 24- 480. It appears that the matrix version scales better with increasing number of data compared to columnwise LASSO.

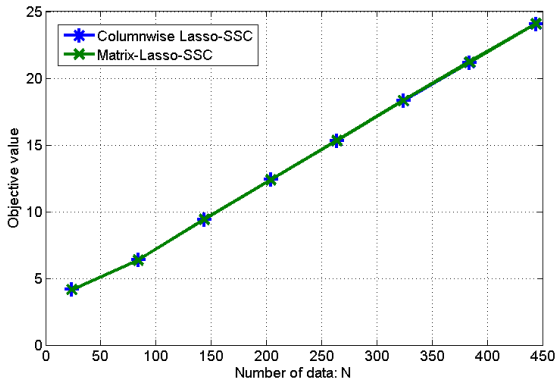


Figure 15: Objective value comparison with increasing number of data. Simulated with $n = 100, d = 4, L = 3, \sigma = 0.2$, κ increases from 2 to 40 such that the number of data goes from 24- 480. The objective value obtained at stop points of two algorithms are nearly the same.

.

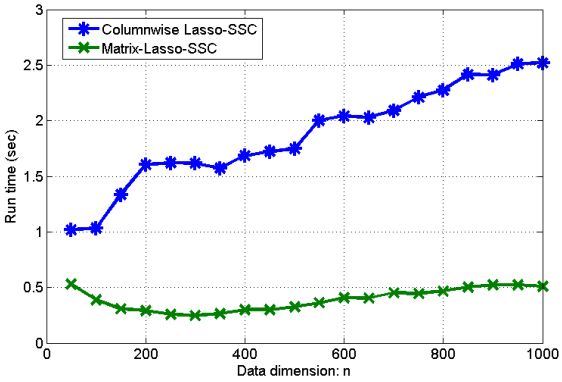


Figure 16: Run time comparison with increasing number of data. Simulated with $\kappa = 5, d = 4, L = 3, \sigma = 0.2$, ambient dimension n increases from 50 to 1000. Note that the dependence on dimension is weak at the scale due to the fast vectorized computation. Nevertheless, it is clear that the matrix version of SSC runs faster.

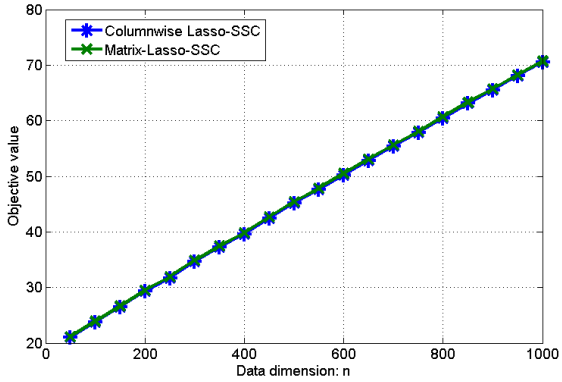


Figure 17: Objective value comparison with increasing number of data. Simulated with $\kappa = 5, d = 4, L = 3, \sigma = 0.2$, ambient dimension n increases from 50 to 1000. The objective value obtained at stop points of two algorithms are nearly the same.

René Brandenberg, Abhi Dattasharma, Peter Gritzmann, and David Larman. Isoradial bodies. *Discrete & Computational Geometry*, 32(4):447–457, 2004.

E.J. Candès. The restricted isometry property and its implications for compressed sensing. *Comptes Rendus Mathematique*, 346(9):589–592, 2008.

G. Chen and G. Lerman. Spectral curvature clustering (scc). *International Journal of Computer Vision*, 81(3):317–330, 2009.

Joao Costeira and Takeo Kanade. *A multi-body factorization method for motion analysis*. Springer, 2000.

J.P. Costeira and T. Kanade. A multibody factorization method for independently moving objects. *International Journal of Computer Vision*, 29(3):159–179, 1998.

S. Dasgupta and A. Gupta. An elementary proof of a theorem of johnson and lindenstrauss. *Random Structures & Algorithms*, 22(1):60–65, 2002.

D.L. Donoho. De-noising by soft-thresholding. *Information Theory, IEEE Transactions on*, 41(3):613–627, 1995.

D.L. Donoho, M. Elad, and V.N. Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *Information Theory, IEEE Transactions on*, 52(1): 6–18, 2006.

Eva L Dyer, Aswin C Sankaranarayanan, and Richard G Baraniuk. Greedy feature selection for subspace clustering. *The Journal of Machine Learning Research*, 14(1):2487–2517, 2013.

E. Elhamifar and R. Vidal. Sparse subspace clustering. In *CVPR'09*, pages 2790–2797. IEEE, 2009.

E. Elhamifar and R. Vidal. Clustering disjoint subspaces via sparse representation. In *ICASSP'11*, pages 1926–1929. IEEE, 2010.

E. Elhamifar and R. Vidal. Sparse subspace clustering: Algorithm, theory, and applications. *arXiv preprint arXiv:1203.1005*, 2012.

Ehsan Elhamifar and Rene Vidal. Sparse subspace clustering: Algorithm, theory, and applications. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(11): 2765–2781, 2013.

B. Eriksson, L. Balzano, and R. Nowak. High rank matrix completion. In *AI Stats'12*, 2012.

T. Hastie and P.Y. Simard. Metrics and models for handwritten character recognition. *Statistical Science*, pages 54–65, 1998.

Reinhard Heckel and Helmut Bölcskei. Noisy subspace clustering via thresholding. In *Information Theory Proceedings (ISIT), 2013 IEEE International Symposium on*, pages 1382–1386. IEEE, 2013.

A. Jalali, Y. Chen, S. Sanghavi, and H. Xu. Clustering partially observed graphs via convex optimization. In *ICML'11*, pages 1001–1008. ACM, 2011.

K. Kanatani. Motion segmentation by subspace separation and model selection. In *ICCV'01*, volume 2, pages 586–591. IEEE, 2001.

Fabien Lauer and Christoph Schnorr. Spectral clustering of linear subspaces for motion segmentation. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 678–685. IEEE, 2009.

Kuang-Chih Lee, Jeffrey Ho, and David J Kriegman. Acquiring linear subspaces for face recognition under variable lighting. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(5):684–698, 2005.

G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma. Robust recovery of subspace structures by low-rank representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):171–184, 2013.

Guangcan Liu, Zhouchen Lin, and Yong Yu. Robust subspace segmentation by low-rank representation. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 663–670, 2010.

Guangcan Liu, Huan Xu, and Shuicheng Yan. Exact subspace segmentation and outlier detection by low-rank representation. In *International Conference on Artificial Intelligence and Statistics*, pages 703–711, 2012.

B. Nasihatkon and R. Hartley. Graph connectivity in sparse subspace clustering. In *CVPR'11*, pages 2137–2144. IEEE, 2011.

A.Y. Ng, M.I. Jordan, Y. Weiss, et al. On spectral clustering: Analysis and an algorithm. In *NIPS'02*, volume 2, pages 849–856, 2002.

Shankar R Rao, Roberto Tron, René Vidal, and Yi Ma. Motion segmentation via robust subspace separation in the presence of outlying, incomplete, or corrupted trajectories. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

M. Soltanolkotabi and E.J. Candes. A geometric analysis of subspace clustering with outliers. *To appear in Annals of Statistics*, 2012.

Mahdi Soltanolkotabi, Ehsan Elhamifar, Emmanuel J Candes, et al. Robust subspace clustering. *The Annals of Statistics*, 42(2):669–699, 2014.

Jos F Sturm. Using sedumi 1.02, a matlab toolbox for optimization over symmetric cones. *Optimization methods and software*, 11(1-4):625–653, 1999.

Ryan J Tibshirani et al. The lasso problem and uniqueness. *Electronic Journal of Statistics*, 7:1456–1490, 2013.

Kim-Chuan Toh, Michael J Todd, and Reha H Tütüncü. Sdpt3a matlab software package for semidefinite programming, version 1.3. *Optimization methods and software*, 11(1-4): 545–581, 1999.

R. Tron and R. Vidal. A benchmark for the comparison of 3-d motion segmentation algorithms. In *CVPR'07*, pages 1–8. IEEE, 2007.

P Tseng. Nearest q-flat to m points. *Journal of Optimization Theory and Applications*, 105 (1):249–252, 2000.

R. Vidal. Subspace clustering. *Signal Processing Magazine, IEEE*, 28(2):52–68, 2011.

R. Vidal, Y. Ma, and S. Sastry. Generalized principal component analysis (gpca). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(12):1945–1959, 2005.

René Vidal, Stefano Soatto, Yi Ma, and Shankar Sastry. An algebraic geometric approach to the identification of a class of linear hybrid systems. In *Decision and Control, 2003. Proceedings. 42nd IEEE Conference on*, volume 1, pages 167–172. IEEE, 2003.

Yu-Xiang Wang and Huan Xu. Noisy sparse subspace clustering. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pages 89–97, 2013.

Yu-Xiang Wang, Huan Xu, and Chenlei Leng. Provable subspace clustering: When lrr meets ssc. In *Advances in Neural Information Processing Systems*, pages 64–72, 2013.

Jingyu Yan and Marc Pollefeys. A general framework for motion segmentation: Independent, articulated, rigid, non-rigid, degenerate and non-degenerate. In *Computer Vision–ECCV 2006*, pages 94–106. Springer, 2006.

A. Zhang, N. Fawaz, S. Ioannidis, and A. Montanari. Guess who rated this movie: Identifying users through subspace clustering. *arXiv preprint arXiv:1208.1544*, 2012.

S.K. Zhou, G. Aggarwal, R. Chellappa, and D.W. Jacobs. Appearance characterization of linear lambertian objects, generalized photometric stereo, and illumination-invariant face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(2): 230–245, 2007.